THE EFFECT ON STATISTICAL INFERENCE OF THE DEGREE OF PRECISION OF RECORDED DATA

by

ANTHONY TRICKER BSc MSc

A thesis submitted to the Council for National Academic Awards in fulfilment of the requirements for the degree of Doctor of Philosophy

Sponsoring Establishment: Department of Applied Statistics and Operational Research
Sheffield City Polytechnic

July 1988

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr GK Kanji and Dr D Preece, under whose supervision and constant encouragement this work has been carried out.

I would also like to thank Dr G Morgan for his support through numerous constructive and stimulating discussions on the subject. I am also grateful to Dr DN Shanbhag and Dr A Allen for their assistance on specific points.

I wish to thank Maggie Bedingham for the excellent typing of this thesis.

Finally I would like to express my gratitude to my wife, Maude, and my children, Edward and Joanna, for all their support.

THE EFFECT ON STATISTICAL INFERENCE OF THE DEGREE OF PRECISION OF ROUNDED DATA

A R TRICKER

ABSTRACT

This thesis concerns the effect of rounding on statistical procedures, where rounding is taken to be the grouping of data at the midpoints of equally spaced intervals.

The characteristic function of the rounded distribution is obtained. This is used to derive general expressions for the moments of univariate and bivariate distributions that have been subject to rounding. The interactive effect of rounding and skewness on the moments is examined.

The performance of certain normal test statistics is examined for rounded data. A study is carried out to obtain precise values for the significance level and power of these statistical tests for rounded data, over many distributions. Guidance is given on what is an appropriate degree of precision for normal data. Special consideration is given to how much non-normality can be allowed without the effect of rounding seriously distorting the significance level and power of a test.

Standard methods of estimating the parameters of a distribution are compared with respect to loss in information caused by rounding. Normal, gamma and exponential distributions are examined. Computational methods are presented for computing the maximum likelihood estimates from rounded normal and gamma data.

In general it is concluded that the effect of rounding on statistical procedures can be increased by the departure from normality of the population. It was found that less precision is required of the recorded data than that which is usually given.

CONTENTS

CHAPTER 1

INTRODUCTION, NOTATION AND REVIEW OF LITERATURE

1.1 Introduction

1.2 Rounding Process, Notation and Terminology

1.3 Literature Review

- 1.3.1 Relationship between the moments of the rounded variable X_R and the underlying continuous variable X
- 1.3.2 Point Estimation
- 1.3.3 Regression
- 1.3.4 Tests of Significance and Confidence Intervals
- 1.3.5 Rules of Rounding

1.1 Introduction

In statistics the rounding of data, ie the grouping of continuous data at the midpoints of equi–spaced intervals, is common.

When data are collected, values are usually rounded to a common degree of precision. This may result from limitations in the accuracy of available measuring devices or cost restrictions necessitating the need for cheap and consequently inaccurate methods of data collection. In other situations rounding may be desirable to simplify subsequent statistical calculations.

As a consequence of rounding, each recorded value will have an associated error, the size of which can have an important effect on the validity of statistical techniques. It therefore becomes important to ensure that the advantages of rounding are not outweighed by distortion in the information obtained.

This thesis investigates the effect on statistical techniques of the degree of precision of the rounded data. There are four main parts in this thesis: a literature review (Chapter 1), a discussion of the implications of rounding on the moments of a distribution (Chapter 2), a study of the behaviour of test statistics for rounded data (Chapters 3, 4 and 5) and a comparison of estimation procedures when the data are rounded (Chapters 6 and 7).

1.2 The Rounding Process, Notation and Terminology

If the values of a continuous random variable X are rounded, the result is a new discrete random variable X_R. If values of X are rounded into intervals of width

w, with midpoints X_R and the centre of the interval containing zero is cw, then X_R has the following values.

$$\ldots\ cw - 3w,\ cw - 2w,\ cw - w,\ cw,\ cw + w,\ cw + 2w,\ \ldots \qquad (1.2\text{–}1)$$

(1.2–1) will be known as the rounding lattice. Here c determines the position of the rounding lattice relative to the origin (zero) and may be located anywhere between $-w/2$ and $w/2$. The mathematical relationship between X and X_R is such that if

$$cw + (n-\tfrac{1}{2})w \leq X < cw + (n+\tfrac{1}{2})w$$
$$\text{then midpoints} \quad X_R = (c+n)w \qquad n = 0, \pm1, \pm2, \ldots \qquad (1.2\text{–}2)$$

The values of X_R will be termed the rounded data.

The following notation will be adopted throughout the thesis for the moments and related measures.

The mth moment of the random variable X (or of the distribution of X) about the origin and mean will be denoted by μ'_m and μ_m respectively.

$$\mu'_m = E[X^m] \quad , \quad \mu_m = E[(X-\mu'_1)^m]$$

Similarly for the random variable X_R (or of the distribution of X_R) we have

$$\mu'_{mR} = E[X_R^m] \quad , \quad \mu_{mR} = E[(X-\mu'_{1R})^m]$$

The following related measures will be used

measure	X	X_R
mean	μ	μ_R
variance	σ^2	$\sigma^2{}_R$
skewness	$\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$	$\sqrt{\beta_{1R}} = \frac{\mu_{3R}}{\sigma_R^3}$
kurtosis	$\beta_2 = \frac{\mu_4}{\sigma_4}$	$\beta_{2R} = \frac{\mu_{4R}}{\sigma_R^4}$

Under the 'usual' decimal rounding rule, which stipulates that data be truncated to a certain number of decimal places (cf Eisenhart 1947) we have c = 0. The rounding lattice is then of the form

$k_1(10)^{k_2}$ where $k_1 = 0, \pm1, \pm2, \ldots$

k_2 is a fixed integer

However this may not always be so, as the following examples show.

0 -	this grouping implies	2.5 -	this grouping implies
2 -	w = 2, c = $\frac{1}{2}$	5.5 -	w = 3, c = $^1/_3$
4 -		8.5 -	
6 -		11.5 -	

(For example: 0 - means, includes all values from zero to less than 2).

When data from a population distribution are rounded, the severity of rounding does not depend solely on w but also on the standard deviation σ of the population. The severity of rounding of a distribution can be regarded as the loss of information due to the process of rounding, which is determined by the number of points on the rounding lattice determining the distribution. If a distribution has a small σ, it will need a smaller w to be represented by a given number of points on the lattice, than does a distribution with a large σ. Hence the loss of information caused by rounding is better measured by the ratio $r = w/\sigma$. This will be called the degree of precision of the rounded data.

Estimation

Throughout the thesis the following terms will be applied to an estimator $\hat{\theta}$ of a parameter θ.

(i) Unbiased estimator

The estimator $\hat{\theta}$ is said to be unbiased for θ if $E[\hat{\theta}] = \theta$. The bias in $\hat{\theta}$ will simply be $E[\hat{\theta}] - \theta$.

(ii) Mean Square Error (MSE)

If $\hat{\theta}$ is an estimator of θ, then the MSE of $\hat{\theta}$ is:

$$MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = V(\hat{\theta}) + [E(\hat{\theta})-\theta]^2.$$

(iii) Efficiency

If $\hat{\theta}_1$, $\hat{\theta}_2$ are estimators of θ based on the same sample size, then the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is the ratio

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}$$

Summary of notation used in the rounding process

w	–	width of the rounding interval
c	–	position of the rounding lattice relative to origin
r	–	the degree of precision of the rounded data
X	–	continuous random variable
X_R	–	a discrete random variable obtained when X is rounded into intervals of width w, with lattice position c, and takes the value of the interval midpoints
μ	–	mean of X
μ_R	–	mean of X_R
σ^2	–	variance of X
σ^2_R	–	variance of X_R
$\sqrt{\beta_1}$	–	measure of skewness of X
$\sqrt{\beta_{1R}}$	–	measure of skewness of X_R
β_2	–	measure of kurtosis of X
β_{2R}	–	measure of kurtosis of X_R

1.3 Literature Review

There is a large amount of literature on the grouping of data. There are many different methods of grouping, of which rounding is one. In this literature review the effect of rounding on statistical techniques will be reviewed. Previous work on rounding can be broadly divided into the following areas.

1.3.1 Relationship between the moments of the rounded variable X_R and the underlying continuous variable X

This section concerns the relationship between the moments of the unrounded and rounded variables X and X_R respectively. Much of the past literature has considered this relationship in terms of using Sheppard's corrections and the justification on various types of distributions.

The early work concerning rounding concentrated on the estimation of population moments from grouped data. Sheppard (1898) was the first to establish that for data which had been classified into equally spaced groups, the class centre may be used to calculate the various moments and the bias introduced by this procedure can be connected by the use of Sheppard's corrections. Sheppard sought to find formulae relating the moments of the grouped variables to those of the underlying continuous data distribution. He considered the relation between the mth population moment

$$\mu'_m = E[X^m] = \int_{-\infty}^{+\infty} x^m \, f(x) \, dx$$

of continuous random variable X with p.d.f. f(x) and the mth moment of the grouped data:

$$\mu'_{mR} = \sum_{i=-\infty}^{+\infty} X_{Ri}^{m} \int_{X_{Ri} - w/2}^{X_{Ri} + w/2} f(x)\, dt$$

where the values of X have been grouped into classes of width w and midpoints X_{Ri}. We thus have the very familiar Sheppard's corrections:

$$\mu'_{1R} = \mu'_1$$

$$\mu'_{2R} = \mu'_2 + \frac{w^2}{12}$$

$$\mu'_{3R} = \mu'_3 + \frac{w^2}{4}\mu'_1$$

. .
. .
. .

The general expression for these formulae is

$$\mu'_{mR} = \sum_{j=0}^{\left[\frac{m}{2}\right]} \left[\frac{w}{2}\right]^{2j} \binom{m}{2j} \mu_{m-2j}\, (2j+1)^{-1} \qquad (1.3\text{-}1)$$

where $\left[\frac{m}{2}\right]$ is the integral part of $\frac{m}{2}$.

Sheppard based his corrections on the Euler–Maclaurin sum theorem, which connects summation with integration. In his paper he implied that the conditions for the applicability of the corrections to hold for a given case are

(i) f(x) is continuous on a finite range (a,b)

(ii) f(x) is such that f(x) and the derivatives of f(x) vanish at the limits, ie it has high order contact.

For several years after the publication of this paper the conditions under which the corrections were valid were subject to argument. In particular, what degree of high order contact should a distribution have before the corrections are valid?

Kendall (1938) attempted to clarify when the corrections are valid. He showed that the conditions of the validity of Sheppard's corrections are in the main the conditions under which the remainder term in the Euler–Maclaurin expansion used to derive the corrections may be neglected to a satisfactory degree of approximation. By establishing conditions under which the remainder term is small enough to be neglected, Kendall gave the following statement concerning the validity of Sheppard's corrections.

Let f(x) be a continuous distribution. The Sheppard's corrections (1.3–1) will be accurate to order w^k ie to the order of the terms applied in the corrections, if

(a) the range of f(x), (a,b) is finite

(b) the order of terminal contact is k ie f(x) and its first k derivatives of f(x) vanish at a and b

(c) $\frac{d^{k+1}}{dx^{k+1}}$ [F(x)] is not large in the range (a,b), where

$$F(x) = X^m \int_{-w/2}^{w/2} f(x+\epsilon)d\epsilon$$

(d) $2\left[\frac{m}{2}\right] < k.$

Kendall pointed out that where the range is infinite no set of conditions as given above can be stated. However, many such distributions taper off strongly to zero at the ends of the range in such a way that the corrections will be valid. He drew attention to two situations where the corrections may be inaccurate. These are:

(i) the distribution of f(x) is markedly skewed

(ii) the degree of precision r (w/σ) is large.

Baten (1931) and Wold (1934) derived Sheppard's corrections for the bivariate situation. Hartley (1950) presented a simplified form of Sheppard's correction formulae which are computationally more convenient. Also his corrections in some cases allow for simultaneous correction for the rounding interval and shift of origin.

Expression (1.3–1) is complicated, particularly when inverted so as to express the moments μ'_m as linear functions of the rounded moments μ'_{mR}. When we deal with cumulants the relationship becomes much simpler. Under the same conditions as for the validity of Sheppard's corrections we have

$$K_s = K_{sR} - B_s \frac{w^s}{s} \qquad s > 1 \qquad (1.3\text{-}2)$$

where K_s and K_{sR} are the sth cumulant of X and X_R respectively, and B_s is the sth Bernoulli number.

The alternative result (1.3–2) was proved by Langdon and Ore (1930). Wold (1934) provided a similar result for bivariate distributions.

Corrections to the moments where the distribution does not have terminal contact, as in J and U shapped distributions, were first given by Pearson (1902). Pearson, like Sheppard (1898), based his corrections on the Euler–Maclaurin theorem. Let f(x) be a continuous distribution with finite range (a,b), where the values of x have been grouped into classes of width w. By assuming that a = 0 and that the derivatives of f(x) vanish at b, Pearson obtained the following corrections:

$$\mu_1' = \mu_{1R}' - \frac{w^4}{720} a_3 + \frac{w^6}{30240} a_5$$

$$\mu_2' = \mu_{2R}' - \frac{w^2}{12} - \frac{w^4}{120} a_2 + \frac{w^6}{30240} a_4$$

where $$a_s = - \left[\frac{d^{s-1}}{dx^{s-1}} f(x)\right]_{x=0} \quad \text{for } s = 2,3,\ldots$$

If there is 'high contact at both ends', then $a_s = 0$, and we have the usual Sheppard's corrections.

Using techniques similar to those of Pearson (1902), Pairman and Pearson (1918) considered corrections to moments where there is no terminal contact at one or both ends of the distribution. However, the corrections they developed for the moments are often tedious to make. Sandon (1924) presented a simplified set of formulae for where the distribution has an exponential curve. Following on the work by Pairman and Pearson (1918), Pearse (1928) considered how the moments should be corrected where the distribution has an infinite range. Martin (1934), using a similar approach to that of Pearse, derived moment corrections for where the lower range of a distribution may be unknown and as a result the width of the first class is not known.

Lewis (1935) suggested corrections that were more reliable than Sheppard's but not so difficult to make as those given by Pairman, Pearson and Pearse. His approach was to estimate the frequency curve in each interval in the rounding lattice by a quadratic density function $f(t) = a + bt + ct^2$. When the constants a, b and c have been obtained for each interval, f(t) is used to determine the moments μ_{mR} of the rounded distribution. However his formulae for μ_{mR} are in general long and complex. Davies and Bruner (1943) developed a correction for the second moment by a similar approach to that of Lewis.

The relationship between the moments of the unrounded and rounded distributions X and X_R respectively, has mainly been obtained by the use of Sheppard's corrections. Although these corrections are only approximate and can be unreliable they have been regarded as the accepted method. Several authors have suggested alternative methods to Sheppard's corrections, but none of these has come into general use, the reaosn being that these alternative methods are generally difficult to use and specific to certain distributions.

In communication engineering the parallel to rounding is the quantization of signals. The theory of quantization has been developed by electrical engineers for signal analysis. Some of the results of this theory can be adapted for use with rounded data. However, this work has been mainly ignored in the statistical literature concerned with rounding until Tricker (1984b) used quantization theory to derive a relationship between the first two monents of the distribution of X and X_R. The content of this paper forms part of the work contained in Chapter 2. A copy of this paper can be found at the end of this thesis.

An important contribution to the theory of quantization is due to Widrow (1956, 1961), who derived the characteristic function of a quantized signal. This is equivalent to the characteristic function of a random variable which has been rounded according to a rounding lattice, with rounding interval of width w and lattice position c. Using the characteristic function, Widrow obtained expressions for the mean and variance of normal rounded data. However they are approximate and restricted to a rounding lattice with c equal to zero. He also considered the joint first moment of quantized signals for a bivariate normal distribution, and gave an approximation to the bias in this moment caused by quantization. The approximation is only suitable for c equal to zero and as shown in Chapter 2 is incorrect.

Watts (1961) generalised the approach by Widrow (1956, 1961) to include quantizers in which scaling (multiplying factors) and shifting (addition and subtraction) are allowed. The probability density functions and characteristic function of a quantized signal for univariate and bivariate distributions are derived. Watts said little about the association between quantization and rounding. In Chapter 2 it will be shown how the results of Watts can be adapted to consider the effect of rounding on the moments of a distribution.

Although statistical literature concerned with rounding has made very little reference to the work of quantization, this is not so for other subject areas. In chemistry, Lowell (1980) extended the work of Widrow (1956, 1961) to find the first two moments of normal rounded data for the more general case of non-zero mean. His results could have been obtained more simply by using the work of Watts (1961). However, Lowell made no reference to Watts' paper.

Several authors have considered the relationship between the moments of X and X_R for a normal population. To date, very little has appeared in the literature concerning this relationship with respect to other populations. Holland (1975) considered the effect of rounding on the moments of non-normal data. His approach was as follows.

Let X be a continuous random variable with p.d.f. f(x). Values of X are rounded to a rounding lattice with interval of width w and lattice position c. The result is the rounded random variable X_R. Holland uses the p.d.f. of X_R given by

$$P[X_R = c + nw) = \int_{c + (n-\frac{1}{2})w}^{c + (n+\frac{1}{2})w} f(x)\, dx \quad , \quad n = 0, \pm 1, \pm 2, \ldots$$

to calculate $E[X_R]$ and $V[X_R]$.

His approach does not provide any explicit expressions for the mean and variance of X_R. He considered only distributions where the distribution of X_R has a closed form or are tabulated in great detail. The two non-normal distributions considered are the exponential and triangular. Although the possible effect of the lattice position is ignored, Holland was the first to mention that the shape of the distribution may be important in determining the effect of rounding on the moments of a distribution. Elsewhere, Tricker (1984a) also derived the p.d.f. of a rounded exponential data and investigated the distortion caused by rounding in the mean and variance. Tricker considered the lattice effect, which he showed to be important in determining the effect of rounding on the moments. The content of this paper forms part of the work contained in Chapters 5 and 7. A copy of this paper can be found at the end of this thesis.

Another paper by Tricker (1984b) uses the theory of quantization to obtain explicit functions for the mean and variance of X_R. It examines the interactive effect of rounding and skewness on the moments of a univariate distribution.

In Chapter 2 it will be shown how the theory of quantization can be applied to the process of rounding. A more detailed examination of the effect of rounding on the moments of a distribution than that given in Tricker (1984b) will be presented.

Average Corrections

There is a distinct type of problem which also leads to Sheppard's corrections (1.3–1) often referred to as 'average corrections'. If the rounding lattice is located at random on the distribution, then Sheppard's corrections hold on average, no matter what form the distribution takes. This result was first given by Abernethy (1933). The parallel result for cumulants can be found in Cornish and Fisher (1937). As pointed out by Kendall (1938), Sheppard's corrections regarded as average corrections require the rounding lattice to be located at random. This condition is not always met in practice. For example, many J and U–shaped distributions naturally begin with an interval starting at zero. The rounding lattice is then not located at random and Sheppard's corrections are not legitimate even on average.

Sheppard's corrections are customarily applied to the moments about the mean of the rounded distribution, namely by omitting the dashes in (1.3–1) and putting $\mu'_{1R} = \mu'_1 = 0$. As noted by Kendall (1938) this is legitimate for the same conditions for which the corrections apply to the moments about zero. However,

for average corrections it is no longer valid to drop the dashes in order to obtain corrections for moments about the mean of the rounded distribution.

Craig (1941) presented a set of expressions for the average correction to the moments about the mean. If the location of the rounding lattice is random then, for example, the average correction for the second moment is given by

$$\sigma^2 = \sigma_R^2 - \frac{w^2}{12} + \sigma_m^2$$

where the variance of X and X_R are respectively σ^2 and σ^2_R, and σ^2_m is the variance of the means of the rounded distribution over all possible rounding lattice positions. It is usually unrealistic to be able to obtain σ^2_m and in practice the above correction has limited use.

1.3.2 Point Estimation

Maximum Likelihood Estimation

The method of obtaining a maximum likelihood estimate (MLE) from grouped data has been extensively studied in the literature. We shall be interested in the situation where the ML method can be applied to rounded data, ie where the group intervals are equal and each interval is represented by its midpoint.

The ML procedure has been used by many researchers to estimate the mean and variance of grouped normal data. Gjeddebaek (1949) presented a method of obtaining the MLEs of μ and σ^2 for a normal distribution, whether or not the

grouping is into equal intervals. However his precedure for obtaining the MLEs is troublesome to use because of its iterative character. Gjeddebaek (1956) considered the efficiency of the MLEs of μ and σ^2 from a sample of normal rounded data, where the sample size is large. He defined the efficiency of these estimates as the mean square error (MSE) of the MLE from unrounded data divided by the MSE of the MLE from rounded data. He gave the asymptotic efficiencies of the MLEs of μ and σ^2 for various degrees of precision r. He showed that the position of the rounding lattice (c) has a negligible effect on the efficiencies of μ and σ^2, if the value of r is less than 2.0 and 1.6 respectively. Gjeddebaek (1957, 1959) considered two types of estimator for the mean and variance of a rounded normal distribution, these being the ML and naive estimators. The naive estimators are the usual estimators of the mean and variance applied to the midpoints of the rounding intervals. Gjeddebaek showed that, when Sheppard's corrections are applied to the naive estimators, they are "practically as efficient" as the ML estimators for $r \leq 2.0$ and $n \leq 100$.

Grundy (1952) presented a method of estimating the MLEs of μ and σ^2 for a truncated normal distribution, when the data have been grouped. The MLEs are obtained by using 'adjusted moments' and the effect of grouping on large sample covariance matrix is discussed. When the group intervals are equal (rounding) the iterative method for finding the MLEs is simpler than that given by Gjeddebaek (1949). Swamy (1960) extended the work of Gjeddebaek (1949) by deriving the MLEs of μ and σ^2 for rounded data from a single and doubly truncated normal distribution. He obtained the MLEs by the same troublesome iterative method as was used by Gjeddebaek (1949).

When parameters are estimated from rounded data, there is a probability that the MLE will not exist. Even though this probability may tend to zero as the sample size increases, Kulldorff (1961) referred to what are conditional ML estimators, with the condition being existence. Kulldorff gave sufficient conditions for the existence, uniqueness, consistency and asymptotic efficiency for ML estimators for grouped data. However the set of sufficient conditions are complicated.

Kulldorff (1961) considered the estimation of the parameters of the normal and exponential distributions. He found the MLEs of μ and σ for rounded normal data. He used an iterative process called the 'Scoring System' to obtain the roots of the likelihood equations. This iterative procedure is less laborious and converges more rapidly than the one used by Gjeddebaek (1949). When estimating μ with σ known, Kulldorff showed that the optimum groups (optimum in the sense of minimum asymptotic variance) are not very far from equidistant. However, when estimating σ with μ known the optimum groups are far from equidistant.

Kulldorff also derived the MLEs for both the scale and location parameters in the exponential distribution, together with their asymptotic variances for rounded data. He studied the MLE of the scale parameter. An approximation to the mean and variance of the MLE of the scale parameter is derived by use of asymptotic expansions. Using these approximations, he showed the extent to which the asymptotic properties of the MLE are satisfactory. For rounded data the results indicate that if the exponential distribution is of the form

$$F(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \qquad x \geq 0, \ \theta > 0$$

then the asymptotic mean and variance of the MLE of θ can be safely used for sample sizes in excess of 24, when $r \leq 2.0$.

Tallis and Young (1962) discussed the ML estimation of parameters for truncated normal, log normal and bivariate normal distributions from rounded data. For each distribution they gave only the ML equations. Algner and Goldberger (1970) investigated the problem of estimating the scale parameter in the Pareto distribution from grouped observations. They showed that, for rounded data, the MLE of the scale parameter has an exact solution.

For rounded data, the MLE must usually be obtained iteratively. Since the use of computers in statistics becomes widespread, iterative methods have been presented for finding the MLEs. Generally the iterative techniques are for grouped data, of which rounding is simply a special case.

Gjeddebaek (1949) used the Newton–Raphson iterative process in finding the MLEs of μ and σ from rounded normal data. Although the iterative procedure was not very simple, he provided tables to assist with the computation. Kulldorff (1961) described an iterative method for finding the MLEs of μ and σ, when one of the two parameters is known. He used the method of scoring, due to Fisher (1925). This method is generally far superior to that used by Gjeddebaek (1949), in terms of simplicity and rate of convergence. Tallis and Young (1962) also suggested the method of scoring for finding the MLEs of μ and σ, since an estimate of the variance–covariance matrix is obtained as part of the computational routine. Swan (1969) used a Newton–Raphson procedure written in ALGOL to obtain MLEs of μ and σ^2 from a normal sample which may be grouped in any way. At the same time approximations to their variances and covariances are obtained. Benn and

Sidebottom (1976) gave a method and a FORTRAN program which can be derived from Fisher's method of scoring for ML estimation of location and scale parameters from grouped data. Wolynetz (1979a) described an algorithm and gave a FORTRAN program which was derived from the expectation–maximisation (EM) algorithm, for grouped and censored normal data. Wolynetz (1979b) gave an algorithm for the normal linear model where the dependent variable is subject to rounding. Schader and Schmid (1984) investigated the performance of the EM, Newton–Raphson, scoring and fixed–point methods for obtaining the MLEs of μ and σ from grouped normal data. One of their main results is that the method of scoring is best both in number of iterations and CPU time. Deken (1983) used the EM algorithm to obtain the MLEs of grouped multivariate normal data. His approach was to approximate the E–step in order to reduce the computer time for the EM algorithm. However the formulae used are complicated.

Approximate Maximum Likelihood Estimation

Consider a random sample of n observations from a p.d.f. $f(x \mid \theta)$, rounded into intervals of width w, with midpoints y_i (i=1,...,n). If we let

$$P(y_i) = \int_{y_i-w/2}^{y_i+w/2} f(x \mid \theta)\, dx$$

then the MLE of θ is obtained by maximising

$$L(\theta) = \prod_{i=1}^{n} P(y_i) \qquad (1.3\text{-}3)$$

However, much work is often needed to obtain this, and it is reasonable to search for an approximate method requiring less effort. Lindley (1950) established a

method for obtaining approximate MLEs for parameters of univariate distributions under equal interval groups. His approach was as follows.

Consider a single group of width w with midpoint y_i. By use of Taylor's theorem we have

$$P(y_i) = wf(y_i) + \frac{w^3}{24} f''(y_i) + O(w^4) \qquad (1.3\text{-}4)$$

where $f(y_i)$ and $f''(y_i)$ are the p.d.f. and second derivative respectively evaluated at y_i. From (1.3–3) and (1.3–4) we have

$$L(\theta) = \prod_{i=1}^{n} \left[wf(y_i) \left[1 + \frac{w^2}{24} \frac{f''(y_i)}{f(y_i)} + O(w^3) \right] \right.$$

and

$$\log L(\theta) = \sum_{i=1}^{n} \left[\log wf(y_i) + \frac{w^2}{24} \frac{f''(y_i)}{f(y_i)} + O(w^3) \right]$$

If θ_0 is the MLE of θ using the midpoints y_i, then by applying the Newton–Raphson method, using θ_0 as the first approximation, and adjusted estimate is given by

$$\hat{\theta} = \theta_0 + \delta$$

where
$$\delta = -\frac{w^2}{24} \left[\frac{\sum_{i=1}^{n} \frac{d}{d\theta} [f''(y_i)/f(y_i)]}{\sum_{i=1}^{n} \frac{d^2}{d\theta^2} [\log f(y_i)]} \right] \qquad (1.3\text{-}5)$$

If the third derivative of $f(x|\theta)$ exists the error in δ is of order $O(w^3)$, where if the fourth derivative exists the error is of order $O(w^4)$. Lindley showed that the approximate MLEs are equivalent to Sheppard corrected moment estimators in the case of rounded normal data. He also derived a formula for the loss in efficiency caused by rounding.

Tallis (1967) presented a method for obtaining the approximate MLE for grouped data. The method is a slight but convenient modification of the method of Lindley. The modification replaces the various terms in Lindley's method by their expectation and δ (1.3-5) now becomes the average bias caused by grouping. He also obtained approximate MLEs for parameters for multivariate distributions under equal grouping and univariate distributions under unequal grouping. Tallis shows that the approximate MLEs for rounded bivariate normal data agree with Sheppard corrections given by Wold (1934). In Don (1981) this result of Tallis is generalised to rounded multivariate normal data.

In the area of probit analysis Tocher (1949) obtained MLEs for grouped data. He derived the equations for an iterative solution to the MLEs of the mean μ and variance σ^2 in the underlying normal distribution of grouped probit data. However the steps in the iteration involve lengthy calculations and make the whole process tedious. Using a similar approach to that of Lindley (1950) he obtained approximate MLEs for μ and σ^2 where the data is rounded. These approximate estimates are found to be equivalent to the Sheppard corrected moment estimates.

Other Methods of Estimation

Although the method of ML has been the most common approach for estimating the parameters of a distribution for rounded data, other methods of estimation have been suggested.

A consistent estimator for rounded data was presented by McNeil (1966). It is computationally simpler than the MLE and is consistent even when the MLE is not. However, it is very difficult to apply for multivariate distributions. With today's computing facilities the computational problems in obtaining the MLE have become less important. As a result the McNeil approach has not become a worthwhile alternative to that of the ML.

Others who have considered alternatives to the ML approach are Yoneda and Uchiyama (1956). They advocated the use of ordered statistics to estimate the mean and variance of coarsely rounded normal data. The method of minimum chi-square was suggested by Hughes (1949) to estimate the variance of rounded bivariate normal data.

Often rounding is ignored and standard estimation procedures are applied to the midpoints of the rounding intervals. This naive method of estimation can lead to misleading inferences. Tricker (1984a) showed this for exponential rounded data. He used the interval midpoints to compute the ML estimator for unrounded data. This naive estimator has bias which does not decrease to zero as the sample size increases. He illustrated how the magnitude of the bias is dependent on the degree of precision of the data and on position of the rounding lattice. For sample sizes exceeding 50, he showed how to compensate for rounding and reduce

the bias in this naive ML estimator.

In past literature, estimation with regard to rounded data has concentrated on the normal distribution. In Chapter 7 the efficiency of various estimation procedures will be investigated when applied to non-normal rounded data.

1.3.3 Regression

Consider the regression model

$$\underline{Y} = \underline{1}\,\beta_0 + \underline{X}\,\underline{\beta}_1 + \underline{E} \qquad (1.3\text{-}6)$$

where $\underline{Y}$ is the (n×1) vector of observations of the dependent variable, $\underline{X}$ is the (n×k) matrix of the values of k independent variables, $\underline{1}$ is the (n×1) vector of ones, $\underline{E}$ is the (n×1) vector of errors assumed to be uncorrelated with zero mean and variance σ^2, and $\underline{\beta}_1 = (\beta_1,\ldots,\beta_k)'$ is the (k×1) vector of regression coefficients to be estimated.

Durbin (1954) considered the problem of error of measurement in the variables of the simple regression model that passes through the origin, ie $\beta_0 = 0$ and $k = 1$ in (1.3-6). The error of measurement can be caused by several factors, of which one can be the process of rounding. Durbin devoted part of his paper to the effect of rounding on the least squares estimate (LSE) of β_1. He stated that, under the traditional assumptions concerning errors of measurement, the LSE of β_1 will be unbiased for rounded data. However, as pointed out by Haitovsky (1973, Ch 6), one of the assumptions concerning errors of measurement is that the errors are uncorrelated with the correct values. He showed that, for rounded data, this

is not so and that Durbin's statement about the bias in LSE of β_1 is in doubt.

Haitovsky (1973) considered the rounding in the simple regression model (k = 1 in 1.3–6). He investigated the bias in the LSEs of β_0 and β_1 for rounded data that takes into account the correlation between the rounding error and the correct value. He derived approximate formulae for the bias in the LSEs of β_0 and β_1 when $\underline{Y}$ and $\underline{X}$ are both subject to rounding. He showed that if the distributions of the independent and dependent variable are symmetrical and unimodal then:

(i) the direction in the bias in the LSEs depends on whether the number of categories into which the independent variable is grouped is larger or smaller than the number of categories into which the dependent variable is grouped

(ii) the bias and loss in efficiency in the LSEs is minimized when the independent and dependent variables are grouped into the same number of categories.

Swindel and Bower (1972) considered the regression model (1.3–6) where the independent variables are subject to rounding. They showed rounding will cause the LSEs of the regression coefficients to be biased, and derived bounds on this bias. However these bounds are dependent on knowing the true value of the regression coefficients and the rounding errors.

Beaton, Rubin and Barone (1976) investigated the effect of rounded data on the regression model (1.3–6). They use computer simulation to recreate the unknown $(\underline{X},\underline{Y})$ by adding a uniform rounding error onto the rounded value $(\underline{X}_R,\underline{Y}_R)$. This simulation process is carried out many times, computing the LSE of $\underline{\beta}$ from $(\underline{X},\underline{Y})$

each time. The main point of this method is that these recreated LSEs are useful in showing the possible disturbances due to rounding. Beaton et al (1976) illustrated their technique on the much analysed Longley (1967) data. They showed how high correlation between the independent variables should be avoided, in order to reduce the effect of rounding. A simple adjustment strategy for improving the LSE of $\underline{\beta}$ is given when the data is rounded. They suggested that, for small rounding intervals an improved LSE of $\underline{\beta}$ can be obtained by adding $w^2/12$ to the diagonals of the sample covariance of $(\underline{X}_R, \underline{Y}_R)$.

Beaton et al (1976) assumed that the rounding error is independent of the rounded value and uniformly distributed with mean zero and variance $w^2/12$. In fact the rounding error is conditional on the value of the rounded observation and knowledge of the rounded value can convey information about the rounding error.

Dempster and Rubin (1982) pointed out the failure of Beaton et al to use a conditional distribution for the rounding error. They illustrate by using an artificial example how the adjustment suggested in their paper to compensate for rounding error can lead to increased bias in the LSESs. For model (1.3–6) Dempster and Rubin considered three sample approaches to the problem of rounding in $(\underline{X}, \underline{Y})$ for least squares regression, these being:

(i) ignore the effect of rounding on the data

(ii) add an adjustment of $w^2/12$ to the diagonals of the covariance matrix (Beaton et al; adjustment)

(iii) subtract an adjustment of $w^2/12$ from the diagonals of the covariance matrix (Sheppard's correction).

Approaches (i) and (ii) are shown to lead to considerable bias in the regression coefficients, especially when the design matrix is ill-conditioned.

Likelihood analysis, which uses the conditional distribution of the unrounded values given the rounded value, can produce a more reliable adjustment. When the width of the rounding interval is small, this adjustment is shown to be (iii) above in two situations

(a) $(\underline{X},\underline{Y})$ is jointly normal

(b) normal residuals, large sample and the distribution of the independent variables are 'regular' (meaning the distribution is relatively smooth).

An adjustment (iii) is suitable only for small rounding intervals. For 'moderate' rounding the adjustment required will vary considerably depending on the distributional form of the independent variables.

Cameron (1987) investigated the effect of rounding on parameter estimation in simple regression, where the error has different distributional forms. He compared two methods of estimating the parameters. In the first, the rounding is ignored and the LS estimators are applied to the midpoints of the rounding intervals; this method is referred to as the OLS method. The second is the ML method for grouped data, which recognize the rounding in the data. A simulation experiment demonstrated that, where the errors are normally distributed, the OLS method yields virtually identical point estimators of the intercept and slope compared to those obtained from the grouped ML method, when the degree of precision r is less than one. While for the same range of r, the OLS method gives only slightly different estimates of the error variance. However, when the distribution of errors

deviates from the normal curve, the OLS method does not necessarily perform well. As the skewness in the errors increases, so does the distortion in the parameters when the OLS method is used.

The use of the ML estimators for regression models, for grouped data has lead to some research into iterative procedures. Burridge (1981) determined for a class of regression models (eg normal, logistic, extreme value), where the dependent variable is grouped, a parameterization in which the log-likelihood is guaranteed to be concave, thereby ensuring the existence of a global maximum. He demonstrated by simulation that this reparameterization of the regression model improved considerably the speed of convergence of iterative procedures based on the Newton-Raphson method.

Burridge (1982) presented a more general set of results on concavity of log-likelihoods, derived from grouped data. In general, concavity of the log-likelihood alone does not imply that the MLE exists. In Burridge (1986), for regression models considered in Burridge (1981), a necessary and sufficient condition is given for the existence of MLEs, where data is grouped.

Approximate Maximum Likelihood Estimation in Regression Models

The method of obtaining MLEs of the parameters in a regression model, where the likelihood recognizes the rounded data is often said to produce the "full maximum likelihood" (FML) estimates. As finding these usually calls for considerable computational effort, alternative methods have been used.

Fryer and Pethybridge (1972) extended the approach of Lindley (1950) to obtain approximate MLEs for the parameters in the simple regression model (k = 1 in 1.3–6), when ($\underline{X}$,$\underline{Y}$) has a bivariate normal. They considered having either or both of the variables rounded. Their simulation results suggest that the approximate MLEs and their corresponding variances are adequate substitutes for the FML estimates if the degree of precision r does not exceed 1.6.

Pethybridge (1973) extended the work of Fryer and Pethybridge (1972) to polynomial regression. In Pethybridge (1975) the approximate MLEs in the simple regression model for rounded data are considered from two aspects: firstly their behaviour in small samples, secondly, if slight departures from normality of $\underline{X}$ have any effect on their suitability in large samples. Simulation results suggest that the approximate MLEs and their corresponding variances are suitable substitutes for the FML estimates if r does not exceed 0.8 when the sample size is 25, and if r does not exceed 1.6 for sample sizes of at least 100. Also slight 'non–detectable' departures from normality in $\underline{X}$ are shown not to affect the acceptance of the approximate MLEs as a suitable approximation to the exact MLEs when the sample size is large.

Indrayan and Rustagi (1979) considered the regression model (1.3–6) where the independent variable $\underline{Y}$ is subject to rounding and the errors are normally distributed. They derived approximate MLEs for the parameters in the regression model. Simulation experiments showed how these approximate MLEs compare with those based on unrounded data. However, the authors provided no indication of the value of r that will give a suitable agreement between the approximate MLEs and the MLEs for unrounded data.

1.3.4 Tests of Significance and Confidence Intervals

In the literature the effect of rounding on tests of significance and confidence intervals has been generally unexplored. For example, how will the distribution of a test statistic be altered by rounded data and which statistical tests are sensitive to rounding? Answers to such questions as these have not been well covered in the literature.

William Sealy Gosset, better known as "Student", was probably the first to discuss the effect of rounding on statistical procedures. In his paper on "The probable error of the mean", Student (1908) discussed the statistical effects of using "wide groups" for data. Student's experimental results suggest that the distribution of the one sample t statistic for rounded and unrounded data will be approximately the same if the sample size is large.

A classic statistical procedure for rounded data is the use of Sheppard's corrections. However, Fisher (1936, Ch 3, App D) advised that

> These adjustments should be used for purposes of estimation, but not usually for tests of significance.

This was reiterated by Eisenhart (1947, p203), who pointed out that use of Sheppard's correction can make the t value imaginary as the corrected estimate of variance can be negative. Further reiteration came from Gjeddebaek (1968). He showed that S^2_R/n is a better estimate of the sampling variance of $\bar{X}_R$, than the corrected alternative $(S^2_R - w^2/12)/n$; where $\bar{X}_R$ and S^2_R are the usual estimates of the population mean and variance, applied to rounded data. His advice is that

S^2_R should be used without Sheppard's correction when $\bar{X}_R$ and S^2_R are brought together in testing procedure or in a statment of confidence limits for the population mean.

Krutchoff (1967) pointed out that rounding can cause the F statistic for equality of two variances to have a non-zero probability of a zero in the demoninator. Thus the mean of this statistic will not exist.

Eisenhart (1947) was the first to study in any detail how rounding affects a test statistic. For samples drawn from a rounded normal population, he concluded:

If the sample size n is sufficiently large and the rounding interval width w is sufficiently small to render the discontinuities in the rounded distribution negligible then the distribution of the

(i) test statistic

$$\chi^2 = (n-1)S^2_R/\left[\sigma^2 + \frac{w^2}{12}\right],$$

will closely approximate a χ^2 distribution with n-1 degrees of freedom for rounded data

(ii) test statistic

$$t = \frac{(\bar{X}_R - \mu)}{S_R/\sqrt{n}},$$

will closely approximate a t distribution with n-1 degrees of freedom for rounded data.

Eisenhart was unable to find a rigorous answer to the question of how large the sample size n needs to be for a given w and how small w needs to be for a

given value of n, before (i) and (ii) may be considered correct. His criterion for judging the suitability of a particular coarseness of rounding was based on the probability of the sample variance, S^2_R for the rounded data being zero. He proposed that $P[S^2_R = 0 | n,w] \leq 0.001$ evaluated on the assumption that sampling is from a normal population be adopted as a definition of values of n and w for which (i) and (ii) can be safely used on rounded data. Eisenhart recommended the following combinations of n and r (w/σ) for which the test statistic χ^2 and t can be used to make inferences about σ^2 and μ respectively.

Degree of Precision r	Sample size
$r \leq 0.005$	$n \geq 2$
$r \leq 0.01$	$n \geq 3$
$r \leq 0.5$	$n \geq 5$
$r \leq 0.8$	$n \geq 6$
$r \leq 1.0$	$n \geq 7$

Eisenhart also suggested that, if the values of (n,r) satisfy the recommendation given above, then the F-test for equality of two variances may be applied to rounded data.

The problem with Eisenhart's recommendations is that they are based on the probability of the sample variance from rounded data being zero. This gives no indication of the performance of the test statistic with respect to level of significance or power under rounding. In Chapters 3 and 4 this issue is investigated.

Preece (1982) examined text book examples of the paired t-test with respect to the degree of precision of the recorded data. He illustrated how the paired t statistic depends crucially both on the rounding interval and position of the rounding grid relative to the origin. Using a similar approach to that of Preece, Riley, Bekele and Shrewsbury (1983) investigated the effect of rounding on the analysis of variance. The degree of precision of data sets taken from standard literature was varied, to illustrate how rounding effects the main squares. Their analysis showed that data could be rounded appreciably before loss of information became signficiant.

The investigations by Preece and Riley et al consisted of looking at the effect of rounding on specific examples. As a result no general conclusions could be established about the performance of a test statistic for rounded data.

1.3.5 Rules of Rounding

Throughout the literature various rules have been suggested for the degree of precision that should be used when recording data. The rules are mainly concerned with data rounded from a normal distribution. There seems to be no standard rule. The purpose of this section is to summarise the rounding rules that have been adopted in the literature.

Although Student (1908) discussed the statistical effect of using poor precision in recording data, it was Fisher (1922) who was probably the first to suggest a rule for rounding data. In this paper, Fisher obtains results on the loss in efficiency due to grouping when the parameters of a normal distribution are estimated from a sample which has been rounded. He showed that, provided the degree of precision

r does not exceed a quarter, the loss in efficiency is less than 1 per cent.

Yates (1937) suggested a rule for rounding in terms of significant figures, after working with British agricultural field experiments. He advised that

> only three significant figures need to be retained in a variate for an analysis of variance provided that the standard error of a single observation is not less than 3–5 per cent of the mean (as in the yields of field plots).

Eisenhart (1947) used the work of Fisher (1922) as a basis for his recommendation. He stated that

> the width of the rounding interval [should] be less than one–third, or better, less than one–fourth the standard deviation of random sampling.

Snedecor and Cochran (1967) echoed this

> For accurate work, the advice commonly given is that I [the width of the rounding interval] should not exceed $\sigma/4$.

Cochran and Cox (1957) translated the $\sigma/4$ rule across to experimental data in the words

> the rounding interval should not exceed one quarter of the standard error per observation.

However, Nicholson (1979) wrote of the "often quoted rule of thumb" that the width of the rounding interval should not exceed $\sigma/2$. This was also the advice of Dyke (1974) and Anon (1975).

Although various rules have been suggested for the degree of precision that should be used in recording data, opinion seems to be divided on which rule to use. Riley et al (1983), when investigating the effect of rounding in the analysis of variance, found Yates' (1937) rule too conservative, whereas Dyke's (1974) rule gave a safe degree of precision for every set of data analysed. In Chapters 3–5 it is shown that, for certain normal test statistics, the rounding rules given in the literature are generally too conservative.

CHAPTER 2

EFFECTS OF ROUNDING ON THE MOMENTS OF A PROBABILITY DISTRIBUTION

2.1 Introduction

This chapter considers the effect of the rounding process on the moments of a continuous distribution. This problem spans the modern era of statistics. Information is obviously lost through rounding and the moments of the distribution may be distorted. This chapter is concerned with the relationship between the moments of the unrounded and rounded distribution X and X_R respectively. We will not be concerned with the problem of estimating the moments from rounded data. A relationship between the moments of X and X_R can be obtained by the use of Sheppard's corrections. Over the years these have been universally regarded as the 'acceptable' method. However, as pointed out by Kendall (1938),

> Sheppard's corrections can be dangerous to use if:
>
> (a) the distribution is markedly skew
>
> (b) the distribution is highly concentrated, ie the standard deviation is of the order of the group interval.

As mentioned in the literature review, other researchers have found corrections to the moments where Sheppard's corrections are invalid. However, these corrections are often tedious to make and specific to certain distributions.

This chapter shows how the relationship between the moments of X and X_R can be obtained from the characteristic function of X_R. General expressions for the moments of X_R are derived. These are found to be more reliable than Sheppard's corrections and provide a means for taking the lattice position into account. For the first time the implications of the shape of the distribution on the effect of rounding on the moments will be investigated. Section (2.2) considers the

univariate distribution, while section (2.3) the bivariate distribution.

In communication engineering the parallel to rounding is the quantization of signals. This work to date has been mainly ignored in statistical literature on rounding. Some of the results of quantization theory can be adapted for use with rounded data. Throughout the chapter such results are indicated and used where possible.

2.2 Univariate Distributions

This section deals with the implications of rounding on the moments of a univariate distribution. The approach is via the characteristic function (CF) of the rounded distribution X_R. In communication engineering Widrow (1961) was the first to demonstrate how the CF of quantized signals was the same as the CF of X_R under certain conditions. However his expression for the CF of X_R was suitable only for lattice position c = 0. Watts (1961) extended Widrow's work. He obtains the CF of a quantized signal which has been subject to scaling and shifting. Watts however made very little connection between his work and that of rounding. Section (2.2.1) shows how we can derive the CF of X_R for any r and c from Watts' result. A proof is given for the CF of X_R which is much simpler and more elegant than the one presented by Watts for the CF of quantized signals.

In section (2.2.1) explicit expressions for moments of X_R are obtained for the first time. These are general results, that allow the amount of distortion caused by rounding on the moments to be measured for a given degree of precision r and lattice position c. This change in moments as a result of rounding is shown for the normal and gamma distributions.

To the author's knowledge the only results on the effect of rounding on the moments in non-normal distributions are by Holland (1975). He considers the first two moments of the rounded exponential and triangular distributions. However, he considered only the first two moments of these distributions. Also he ignored the possible effect of lattice position. In order to gain some insight into the behaviour of the moments from rounded non-normal data, a system of distributions is looked at. For each distribution in the system the effect of rounding on the first four moments about the mean is investigated, for various degrees of precision and lattice positions.

2.2.1 Characteristic Function and Moments of X_R

Throughout this chapter the characteristic function (CF) $\varphi_X(t)$ of a random variable X with p.d.f. f(x) is defined by

$$\varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)\, dx$$

Theorem 2.1

Let X be a continuous random variable with p.d.f. f(x) and CF $\varphi_X(t)$. Values of X are rounded, corresponding to a rounding lattice with intervals of width w and lattice position c. The result is the rounded random variable X_R. Under slight regularity conditions for f(x), the CF of X_R is given by

$$\varphi_{XR}(t) = \sum_{k=-\infty}^{+\infty} e^{-i\,2\pi kc}\ \varphi_X\left[t + \frac{2\pi k}{w}\right]\ \frac{\mathrm{Sin}\frac{1}{2}(tw+2\pi k)}{\frac{1}{2}(tw+2\pi k)} \qquad (2.2\text{-}1)$$

For the proof of this theorem we require the following results.

Fourier Transform (FT)

Let g(x) be a continous function defined on $-\infty < x < \infty$. Then the FT of g(x) is

$$F_g(u) = \int_{-\infty}^{+\infty} e^{-iux} g(x)\ dx$$

Note: $\varphi_g(-u) = F_g(u)$ where g has unit mass.

Convolution Theorem

Let $f_1(x)$ and $f_2(x)$ be two given functions, with FTs $F_1(u)$ and $F_2(u)$ respectively. The convolution of $f_1(x)$ and $f_2(x)$ is given by

$$f(x) = \int_{-\infty}^{+\infty} f_1(s)\ f_2(x-s)\ ds$$

The FT of f(x) is $F_1(u)F_2(u)$.

The Poisson Summation Formula (eg Dym and Mckean, 1972)

If $g(x)$ has FT $F_g(u)$ then for each fixed T,

$$\sum_{n=-\infty}^{+\infty} g(x-nT) = \frac{1}{T} \sum_{n=-\infty}^{+\infty} e^{\frac{in2\pi x}{T}} F_g\left[\frac{2\pi n}{T}\right]$$

provided g is twice differential and satisfying additionally

$$|g(x)| + |g'(x)| + |g''(x)| \leq \text{constant} \times (1+x^2)^{-1}$$

Proof Theorem 2.1

Let the random variable X have p.d.f. f(x). This distribution is rounded into intervals of width w and the centre of the initial interval is a.

Consider the function

$$P(a) = \int_{-\frac{w}{2}+a}^{\frac{w}{2}+a} f(x)\, dx = \int_{-\frac{w}{2}}^{\frac{w}{2}} f(a-x)\, dx$$

Let $k(x) = 1$ for $|x| \leq \frac{w}{2}$

$= 0$ elsewhere

Then $P(a) = \int_{-\infty}^{+\infty} k(x)\, f(a-x)\, dx$ which is a convolution and the FT of P(a) using the convolution theorem is:

$$F_p(u) = F_k(u)F_f(u) = \frac{\sin\left[\frac{uw}{2}\right]}{\frac{u}{2}} F_f(u) \qquad (1)$$

Consider a new function $G(a) = e^{ita}P(a)$. Then the CF of the rounded random variable is

$$\varphi_{XR}(t) = \sum_{n=-\infty}^{+\infty} e^{it(a+nw)} P(a+nw) = \sum_{n=-\infty}^{+\infty} G(a+nw) \qquad (2)$$

Hence provided G(a) is twice differentiable and bounded as indicated, the Poisson Summation Formula gives

$$\varphi_{XR}(t) = \frac{1}{w} \sum_{n=-\infty}^{+\infty} e^{\frac{i\,2\pi na}{w}} F_G\left[\frac{2\pi n}{w}\right] \qquad (3)$$

[It is the conditions on G that impose the regularity conditions on f. In general for statistical distributions these conditions will be satisfied.]

However, the FT of G(a) equals the FT of $e^{ita}P(a)$, ie

$$F_G(u) = F_p(u-t) = \frac{\mathrm{Sin}\left[\frac{w}{2}(u-t)\right]}{\frac{1}{2}(u-t)} F_f(u-t) \quad \text{from (1)}$$

Hence from (3)

$$\varphi_{XR}(t) = \frac{1}{w} \sum_{n=-\infty}^{+\infty} e^{\frac{i2\pi na}{w}} F_f\left[\frac{2\pi n}{w} - t\right] \frac{\mathrm{Sin}\left[\frac{w}{2}\left[\frac{2\pi n}{w} - t\right]\right]}{\frac{1}{2}\left[\frac{2\pi n}{w} - t\right]}$$

As f(x) is a p.d.f., $F_f\left[\frac{2\pi n}{w} - t\right] = \varphi_X\left[t - \frac{2\pi n}{w}\right]$ where $\varphi_X(.)$ is the CF of X. Letting k = -n we have

$$\varphi_{XR}(t) = \sum_{k=-\infty}^{+\infty} e^{-\frac{i2\pi ka}{w}} \varphi_X\left[t + \frac{2\pi k}{w}\right] \frac{\mathrm{Sin}\left[\frac{1}{2}(tw+2\pi k)\right]}{\frac{1}{2}(tw+2\pi k)} \qquad (4)$$

Letting the centre of the initial interval be a = cw, where cw is the centre of interval containing zero, we have the required results.

Watts (1961) derived the CF of the general quantizer system for an electrical signal. By letting the gain and shift be equal to w and c respectively in the quantizer system, we can obtain (2.2-1). Although in communiction engineering a similar result to (2.2-1) has been used for the CF of the quantizer output, Tricker (1984) was the first to apply the parallel result for rounded data in the statistical literature. Watts' method of obtaining the CF of the quantized signal was to first find the p.d.f. and then derive the CF from this. A simpler approach could be applied to find the CF of X_R. However the proof given above is far simpler than if we were to use Watts' approach.

The moments of the rounded random variable X_R can be derived from its CF (2.2-1).

Corollary 2.1 to Theorem 2.1

Let X_R be the rounded random variable with CF $\varphi_{XR}(t)$. Assume the nth moment of X exist. X_R has the following moments.

For n odd

$$E\left[X_R^n\right] = \sum_{k=0}^{\left[\frac{n}{2}\right]} {}^nC_{2k}\left[\frac{w}{2}\right]^{2k}(2k+1)^{-1}E\left[X^{n-2k}\right]$$

$$+ 2(-1)^{\frac{n+3}{2}} \sum_{k=1}^{\infty}\sum_{r=1}^{n} {}^nC_r f^r(\pi k)\left[B^{n-r}\mathrm{Cos}(2\pi kc) - A^{n-r}\mathrm{Sin}(2\pi kc)\right] \qquad (2.2\text{-}2)$$

For n even

$$E\left[X_R^n\right] = \sum_{k=0}^{\left[\frac{n}{2}\right]} {}^nC_{2k}\left[\frac{w}{2}\right]^{2k}(2k+1)^{-1}E\left[X^{n-2k}\right]$$

$$+ 2(-1)^{\frac{n}{2}} \sum_{k=1}^{\infty}\sum_{r=1}^{n} {}^nC_r f^r(\pi k)\left[A^{n-r}\mathrm{Cos}(2\pi kc) + B^{n-r}\mathrm{Sin}(2\pi kc)\right] \qquad (2.2\text{-}3)$$

where (i) $\left[\frac{n}{2}\right]$ is the integer part of $\frac{n}{2}$

(ii) $A^o + iB^o = \varphi_X\left[\frac{2\pi k}{w}\right]$

$$A^s + iB^s = \left[\frac{d^s\varphi_X(t)}{dt^s}\right]_{t = \frac{2\pi k}{w}} \qquad s = 1, 2, \ldots$$

(iii) $$f^s(\pi k) = \left[\frac{w}{2}\right]^s \left[\frac{d^s}{dt^s}\left[\frac{\text{Sin } t}{t}\right]\right]_{t = \pi k} \qquad s = 1, 2, \ldots$$

Proof of Corollary 1

The CF of X_R is

$$\varphi_{XR}(t) = \sum_{k=-\infty}^{+\infty} e^{-i2\pi kc} \; \varphi_X\left[t + \frac{2\pi k}{w}\right] \frac{\text{Sin } \frac{1}{2}(tw+2\pi k)}{\frac{1}{2}(tw+2\pi k)} \qquad (1)$$

The moments of X_R are given by

$$E\left[X_R^n\right] = (-i)^n\left[\varphi_{XR}^n(t)\right]_{t=0} \text{ where } \varphi_X^n(t) = \frac{d^n}{dt^n}\left[\varphi_{XR}(t)\right] \qquad (2)$$

$$\varphi_X^n(t) = \frac{d^n}{dt^n}\left[\sum_{k=-\infty}^{+\infty} e^{-i2\pi kc}\varphi_X\left[t + \frac{2\pi k}{w}\right] \frac{\text{Sin } \frac{1}{2}(tw+2\pi k)}{\frac{1}{2}(tw+2\pi k)}\right]$$

Interchanging the order of summation and differentiation

$$\varphi_X^n(t) = \sum_{k=-\infty}^{+\infty} e^{-i2\pi kc} \frac{d^n}{dt^n}\left[\varphi_X\left[t + \frac{2\pi k}{w}\right] \frac{\text{Sin } \frac{1}{2}(tw+2\pi k)}{\frac{1}{2}(tw+2\pi k)}\right]$$

$$= \sum_{k=-\infty}^{+\infty} e^{-i2\pi kc}\left[\sum_{r=0}^{n} {}^nC_r \varphi_X^{n-r} f^r\right] \qquad (3)$$

where $$ {}^{n}C_{r} = \frac{n!}{(n-r)!\,r!} $$

$$ \varphi_X^j = \begin{cases} \varphi_X\left[t + \frac{2\pi k}{w}\right] & j = 0 \\ \frac{d^j}{dt^j}\left[\varphi_X\left[t + \frac{2\pi k}{w}\right]\right] & j = 1,2,\ldots,n \end{cases} $$

$$ f^j = \begin{cases} \frac{\text{Sin } \frac{1}{2}(tw+2\pi k)}{\frac{1}{2}(tw+2\pi k)} & j = 0 \\ \frac{d^j}{dt^j}\left[\frac{\text{Sin } \frac{1}{2}(tw+2\pi k)}{\frac{1}{2}(tw+2\pi k)}\right] & j = 1,2,\ldots,n \end{cases} $$

From (2) and (3)

$$ E\left[X_R^n\right] = (-i)^n\left[\varphi_{XR}^n(t)\right]_{t=0} $$

$$ = (-i)^n \sum_{k=-\infty}^{+\infty} e^{-i\,2\pi kc}\left[\sum_{r=0}^{n} {}^{n}C_r \varphi_X^{n-r}\left[\frac{2\pi k}{w}\right] f^r(\pi k)\right] \qquad (4) $$

where $$ \varphi_X^j(z) = \begin{cases} \varphi_X(z) & j = 0 \\ \left[\frac{d^j}{dt^j}\,\varphi_X(t)\right]_{t=z} & j = 1,\ldots,n \end{cases} $$

$$ f^j(z) = \begin{cases} \frac{\text{Sin } z}{z} & \\ \left[\frac{w}{2}\right]^j\left[\frac{d^j}{dt^j}\left[\frac{\text{Sin } t}{t}\right]\right]_{t=z} & j = 1,\ldots,n \end{cases} $$

From 4

$$E\left[X_R^n\right] = \overbrace{(-i)^n \sum_{r=0}^{n} {}^nC_r \varphi_X^{n-r}(0) f^r(0)}^{S_1}$$

$$\overbrace{+ (-i)^n \sum_{k=1}^{\infty} \sum_{r=0}^{n} {}^nC_r \left[e^{-i2\pi kc} \varphi_X^{n-r}\left[\frac{2\pi k}{w}\right] f^r(\pi k) + e^{i2\pi kc} \varphi_X^{n-r}\left[-\frac{2\pi k}{w}\right] f^r(-\pi k)\right]}^{S_2} \qquad (5)$$

The first summation of (5), S_1 is the corrections to the moments given by Sheppard's corrections.

We have

$$f^r(0) = \begin{cases} \left[\frac{w}{2}\right]^r \left[\frac{1}{r+1}\right](-1)^{r/2} & \text{for } r = 2k \ (k=0,1,\ldots) \\ 0 & \text{for } r = 2k+1 \end{cases}$$

hence

$$S_1 = \sum_{k=0}^{\left[\frac{n}{2}\right]} {}^nC_{2k} \left[\frac{w}{2}\right]^{2k} (2k+1)^{-1} (-1)^k (-i)^n \varphi_X^{n-2k}(0)$$

$$= \sum_{k=0}^{\left[\frac{n}{2}\right]} {}^nC_{2k} \left[\frac{w}{2}\right]^{2k} (2k+1)^{-1} (-i)^{n-2k} \varphi_X^{n-2k}(0)$$

$$= \sum_{k=0}^{\left[\frac{n}{2}\right]} {}^nC_{2k} \left[\frac{w}{2}\right]^{2k} (2k+1)^{-1} E[X^{n-2k}] \qquad (6)$$

(6) is the expression for Sheppard's corrections (eg Kendall & Stuart, 1968).

For the second summation S_2

$$S_2 = (-i)^n \sum_{k=1}^{\infty} \sum_{r=0}^{n} {}^nC_r \left[e^{-i2\pi kc} \varphi_X^{n-r}\left[\frac{2\pi k}{w}\right] f^r(\pi k) + e^{i2\pi kc} \varphi^{n-r}\left[-\frac{2\pi k}{w}\right] f^r(-\pi k)\right]$$

For the function $f^r(.)$, $f^0(\pi k) = 0$, and $f^r(-z) = (-1)^r f(z)$, thus

$$S_2 = (-i)^n \sum_{k=1}^{\infty} \sum_{r=1}^{n} {}^nC_r f^r(\pi k) \left[e^{-i2\pi kc} \varphi_X^{n-r}\left[\frac{2\pi k}{w}\right] + e^{i2\pi kc} \varphi_X^{n-r}\left[-\frac{2\pi k}{w}\right](-1)^r\right]$$

$$= (-i)^n \sum_{k=1}^{\infty} \sum_{r=1}^{n} {}^nC_r f^r(\pi k) \left[\left[\varphi_X^{n-r}\left[\frac{2\pi k}{w}\right] + (-1)^r \varphi_X^{n-r}\left[-\frac{2\pi k}{w}\right]\right] \mathrm{Cos}(2\pi kc) \right.$$

$$\left. - \left[\varphi_X^{n-r}\left[\frac{2\pi k}{w}\right] - (-1)^r \varphi^{n-r}\left[-\frac{2\pi k}{w}\right]\right] i\ \mathrm{Sin}(2\pi kc)\right] \qquad (7)$$

Let

$$\varphi_X^j(t) = \frac{d^j}{dt^j}\left[\varphi_X(t)\right] = \begin{cases} \varphi_X(t) = A^0(t) + iB^0(t) & j = 0 \\ A^j(t) + iB^j(t) & j = 1, \ldots, n \end{cases}$$

then

$$\left.\begin{aligned} \varphi_X^j(t) + \varphi_X^j(-t) &= 2A^j(t) \\ \varphi_X^j(t) - \varphi_X^j(-t) &= i2B^j(t) \end{aligned}\right] \quad \text{for } j = 0, 2, 4, \ldots \qquad (8)$$

$$\left.\begin{aligned} \varphi_X^j(t) + \varphi_X^j(-t) &= i2B^j(t) \\ \varphi_X^j(t) - \varphi_X^j(-t) &= 2A^j(t) \end{aligned}\right] \quad \text{for } j = 1, 3, 5, \ldots$$

The results in (8) can easily be shown to be true, using Corollary 2 (Lukacs, 1970 pp22).

For n odd

Using (7) and the CF results given in (8), the S_2 for n odd denoted by S_0 is

$$S_0 = -2(-i)^{n+1} \sum_{k=1}^{\infty} \sum_{r=1}^{n} {}^{n}C_r f^r(\pi k)\left[B^{n-r}\mathrm{Cos}(2\pi kc) - A^{n-r}\mathrm{Sin}(2\pi kc)\right]$$

$$= 2(-1)^{\frac{n+3}{2}} \sum_{k=1}^{\infty} \sum_{r=1}^{n} {}^{n}C_r f^r(\pi k)\left[B^{n-r}\mathrm{Cos}(2\pi kc) - A^{n-r}\mathrm{Sin}(2\pi kc)\right]$$

where

$$A^0 + iB^0 = \varphi_X\left[\frac{2\pi k}{w}\right]$$

$$A^s + iB^s = \left[\frac{d^s\varphi_X(t)}{dt^s}\right]_{t = \frac{2\pi k}{w}} = \varphi_X^s\left[\frac{2\pi k}{w}\right] \qquad s = 1,\ldots,n$$

$$f^s(\pi k) = \left[\frac{w}{2}\right]^s\left[\frac{d^s}{dt^s}\left[\frac{\mathrm{Sin}\ t}{t}\right]\right]_{t = \pi k} \qquad s = 1,\ldots,n$$

For n even

Using the same approach as for n odd, the summation for n even, denoted by S_E is

$$S_E = 2(-1)^{\frac{n}{2}} \sum_{k=1}^{\infty} \sum_{r=1}^{n} {}^{n}C_r f^r(\pi k)\left[A^{n-r}\mathrm{Cos}(2\pi kc) - B^{n-r}\mathrm{Sin}(2\pi kc)\right]$$

Hence

$$E\left[X_R^n\right] = \begin{cases} S_1 + S_0 & n \text{ odd} \\ S_1 + S_E & n \text{ even} \end{cases}$$

Under general conditions for the distribution of X, an explicit relationship between the moments of unrounded and rounded data has been obtained. This allows us to examine the effect of rounding on the moments of a distribution. In the past Sheppard's corrections have often been used uncritically to obtain a relationship between the moments of X and X_R. Although to some extent Sheppard's corrections may be used for this purpose, they are only approximate. Our results are more general than Sheppard's corrections in that they allow correction for rounding with respect to w and c.

Sheppard's Corrections

$$\begin{aligned}
\mu'_{1R} &= \mu'_1 \\
\mu'_{2R} &= \mu'_2 + \frac{w^2}{12} \\
\mu'_{3R} &= \mu'_3 + \frac{w^2}{4}\mu'_1 \\
\mu'_{4R} &= \mu'_4 + \frac{w^2}{2}\mu'_2 + \frac{w^4}{80} \\
&\vdots \quad\quad \vdots \quad\quad \vdots
\end{aligned} \qquad (2.2\text{-}4)$$

Sheppard's corrections are customarily applied to the moments about the mean of the rounded distribution, namely by omitting the dashes in (2.2–4) and putting $\mu'_{1R} - \mu'_1 = 0$. Using these corrections, a relationship between the mean, variance, skewness and kurtosis of X and X_R can be obtained.

$$\mu_R = \mu$$

$$\sigma_R^2 = \sigma^2\left[1 + \frac{r^2}{12}\right]$$

$$\sqrt{\beta_{1R}} = \sqrt{\beta_1} \Big/ \left[1 + \frac{r^2}{12}\right]^{3/2} \qquad (2.2\text{-}5)$$

$$\beta_{2R} = \left[\beta_2 + \frac{r^2}{2} + \frac{r^4}{80}\right] \Big/ \left[1 + \frac{r^2}{12}\right]^2$$

where r w/σ.

The expressions in (2.2–5) will be valid only if Sheppard's corrections are also valid for the moments about the mean.

For some distributions occurring in practice which have resonably high order contact Sheppard's corrections (2.2–4) will provide reasonable results. However as pointed out in section (2.1), there are two cases in particular where such an assumption would be dangerous, if the distribution is markedly skewed or rounding is coarse. Where there is no high order contact, such as J and U shaped distributions, Sheppard's corrections may break down. Corrections for such situations have been considered by Pairman and Pearson (1918) and Pearse (1928) but are very complicated and tedious to carry out.

If the rounding lattice itself is considered a random quantity, Sheppard's corrections hold on average, no matter what the unrounded distribution. However, the assumption of a random imposition of rounding lattice seems unreasonable in most circumstances. For example a randomly imposed rounding lattice may introduce data rounded to values outside a distribution's range of positive probability.

Derived from the CF of X_R (2.2-2) and (2.2-3) are an alternative to such methods as Sheppard's corrections. They are more reliable than these corrections on a wider class of distributions. Doubt about the validity of Sheppard's corrections can also be demonstrated by examining the CF of X_R. When the condition (2.2-6) is placed on the CF of X, only the central section of (2.2-1) enters the calculation of the moments of X_R.

$$\varphi_X(t) = 0 \qquad |t| > \frac{2\pi}{w}$$

or (2.2-6)

$$\varphi_X\left[\frac{2\pi k}{w}\right] = 0 \qquad k = \pm 1, \pm 2, \ldots$$

ie the CF of X vanishes outside of a finite interval of t.

The expression for the central section of (2.2-1) is

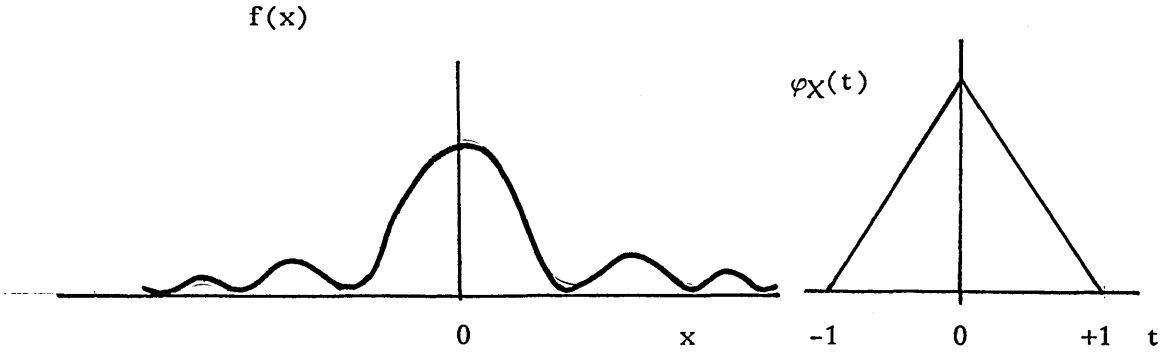
$$\varphi_{X_R}(t)\Big|_{\text{central section}} = \varphi_X(t)\,\frac{\text{Sin}(tw/2)}{tw/2} \qquad (2.2\text{-}7)$$

The central section (2.2-7) can be thought of as a CF in its own right. It is the product of the CF of X and a variable which is uniform, distributed between -w/2 and w/2. Satisfaction of condition (2.2-6) suggests that the moments of X_R are the same as those of the sum of the moments of X and a statistically independent error, uniformly distributed on (-w/2,w/2).

This statement implies that Sheppard's corrections are valid if (2.2-6) holds, ie if the CF of X vanishes outside of a finite interval. Distributions whose CF is zero outside of a finite range of t do exist. An example of such a distribution is given

below.

$$f(x) = \frac{1 - \text{Cos } x}{\pi x^2} \qquad -\infty < x < \infty \qquad \varphi_X(t) = \begin{cases} 1 - |t| & |t| \leq 1 \\ 0 & |t| > 1 \end{cases}$$



However the moments of this distribution do not exist.

Polya's theorem (Lukas, 1970) or the method presented by Kawata (1940) may be used to construct a CF which vanishes outside of a finite interval. However, statistical distributions where the CF has this property are rare. We may deduce that the validity of Sheppard's corrections can be in doubt, as satisfaction of (2.2-6) is uncommon, the reason being that it is rare to have a probability distribution whose CF is zero outside of a finite range of t. But the value of the CF outside of the region (2.2-6) is often very small and may be regarded as negligible for the accuracy we are interested in. As a result in many practical situations Sheppard's corrections may be reliable.

Kullback (1935) gave a proof of Sheppard's corrections based on the CF. He shows that if the distribution has high order contact then

$$\varphi_{X_R}(t) = \frac{2}{wt} \operatorname{Sin} \frac{tw}{2} \varphi_X(t) \qquad (2.2\text{-}8)$$

Clearly from the work in this section, (2.2-8) is only an approximate result. This point is not made clear from his paper.

The next section uses the expressions given in Theorem (2.2) to find the first four moments of the rounded distribution. At the same time the reliability of Sheppard's corrections and the effect of the shape of the distribution on the rounding process are examined.

2.2.2 Moments of rounded Normal and Gamma data

From Corollary 1 the first four moments of X_R are given by

$$E[X_R] = E[X] + \frac{w}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \left[B^0 \operatorname{Cos}[2\pi kc] - A^0 \operatorname{Sin}[2\pi kc]\right] \qquad (2.2\text{-}9)$$

$$E[X_R^2] = E[X^2] + \frac{w^2}{12} - \sum_{k=1}^{\infty} (-1)^k \left[\frac{2w}{\pi k}\left[A' \operatorname{Cos}[2\pi kc] + B' \operatorname{Sin}[2\pi kc]\right] - \left[\frac{w}{\pi k}\right]^2 \left[A^0 \operatorname{Cos}(2\pi kc) + B^0 \operatorname{Sin}[2\pi kc]\right]\right] \qquad (2.2\text{-}10)$$

$$E[X_R^3] = E[X^3] - 3\sum_{k=1}^{\infty}(-1)^k\left[\left[\frac{w}{\pi k}\right]\left[B^2\,\mathrm{Cos}(2\pi kc) - A^2\,\mathrm{Sin}(2\pi kc)\right]\right.$$

$$- \left[\frac{w}{\pi k}\right]^2\left[B'\,\mathrm{Cos}(2\pi kc) - A'\,\mathrm{Sin}(2\pi kc)\right]$$

$$\left.+ \frac{1}{12}\left[\frac{w}{\pi k}\right]^3\left[6-(\pi k)^2\right]\left[B^0\,\mathrm{Cos}(2\pi kc) - A^0\,\mathrm{Sin}(2\pi kc)\right]\right] \qquad (2.2\text{-}11)$$

$$E[X_R^4] = E[X_4] + \frac{w^2}{2}E[X_2] + \frac{w^4}{80}$$

$$+ \sum_{k=1}^{\infty}(-1)^k\left[\left[\frac{4w}{\pi k}\right]\left[A^3\,\mathrm{Cos}(2\pi kc) + B^3\,\mathrm{Sin}(2\pi kc)\right]\right.$$

$$- 6\left[\frac{w}{\pi k}\right]^2\left[A^2\,\mathrm{Cos}(2\pi kc) + B^2\,\mathrm{Sin}(2\pi kc)\right]$$

$$+ \left[\frac{w}{\pi k}\right]^3\left[6 - (\pi k)^2\right]\left[A'\,\mathrm{Cos}(2\pi kc) + B'\,\mathrm{Sin}(2\pi kc)\right]$$

$$\left.+ \frac{1}{2}\left[\frac{w}{\pi k}\right]^4\left[(\pi k)^2-6\right]\left[A^0\,\mathrm{Cos}(2\pi kc) + B^0\,\mathrm{Sin}(2\pi kc)\right]\right] \qquad (2.2\text{-}12)$$

where $A^0 + iB^0 = \varphi_X\left[\frac{2\pi k}{w}\right]$

$$A^s + iB^s = \left[\frac{d^s\varphi_X(t)}{dt^s}\right]_{t = \frac{2\pi k}{w}} \qquad s = 1,2,\ldots$$

Note

(i) The expressions for the moments of X_R simplify considerably if the distribution of X is symmetric (f(x) = f(−x)), as the CF is real and $B^s = 0$ for all s.

(ii) $\mu_{m,R}$ can be obtained by using results connecting the moments about the origin and mean (cf Kendall & Stuart, 1968).

Normal Distribution

As the normal distribution is central to the theory of statistical inference, the effect of rounding on this distribution is important. Several authors have looked at the problem of rounding in the normal distribution with respect to the moments. Of course we have the well known approach of Sheppard's corrections. The first real attempt to use Fourier analysis in the investigation of grouped moments was by Fisher (1922) who expressed the grouped moments in terms of a Fourier Series. He obtained an approximation to the first four moments of the standard normal distribution where the data has been grouped. In fact his approximation takes into account only the first periodic term in the Fourier Series. This is equivalent to using (2.2-9 to 2.2-12) with k = 1. He concludes that if the rounding interval is less than the sd ($r < 1$) the periodic terms are small and may be ignored for normal rounded data. This is rather conservative, as is shown later. Fisher makes no real attempt to investigate the influence of the lattice position or distributions other than normal.

From the CF of X_R, Widrow (1961) obtained expressions for the mean and variance of normal rounded data. They are approximate and equivalent to using (2.2-9) and (2.2-10) with k = 1 and c = 0. As shown later, Widrow's method will give a good approximation to the mean and variance of normal rounded data for $r < 1$. However this may not be so for non-normal rounded data, as disregarding the effect of lattice position will cause Widrow's approximation to be inaccurate. Lowell (1980) gave expressions for the mean and variance of normal

rounded data for the more general case of non-zero mean. In this paper only the first two moments were investigated with respect to rounding.

The expressions for the moments of X_R given in the previous section, will be used to show how the moments in a normal distribution may be affected by rounding. This will extend the work of previous researchers, in that it will consider the effect of both the degree of rounding and of the lattice position on the first four moments and related measures of skewness and kurtosis.

In order to gain insight into the effect of rounding on the moments, the first moment will be considered first.

Let the random variable X have a normal distribution with mean μ and variance σ^2. Let X_R be the random variable corresponding to a rounding lattice with interval of width w and lattice position c. Using the CF

$$\varphi_X(t) = \exp\left[-\frac{t^2\sigma^2}{2} + it\mu\right]$$

we obtain from (2.2-9)

$$E[X_R] = E[X] + \frac{w}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\exp\left[-2\left[\frac{\pi k\sigma}{w}\right]^2\right]\mathrm{Sin}\left[2\pi k\left[\frac{\mu}{w} - c\right]\right] \qquad (2.2\text{-}13)$$

Let μ = cw + nw + αw, where the mean μ lies a distance αw from the nearest midpoint cw + nw and r = w/σ, then from (2.2-13)

$$E[X_R] = E[X] + \sigma\left[\frac{r}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\exp\left[-2\left[\frac{\pi k}{r}\right]^2\right]\mathrm{Sin}(2\pi k\alpha)\right] \quad (2.2\text{-}14)$$

The difference between the expected value of X_R and X will depend on

(i) the ratio $r = w/\sigma$, which is a measure of the severity of rounding

(ii) the value of α, where αw is the distance between the nearest midpoint and the mean μ. α has the range $-\frac{1}{2}$ to $\frac{1}{2}$.

Equation (2.2–14) indicates that the effect of rounding on the first moment of X is dependent not only on the predetermined lattice given by c and w, but also on the position of the normal distribution relative to zero, ie the value of μ. Before the effect of rounding on the moments of the normal distribution can be calculated the values of r, μ and c must be known.

The result (2.2–13), for the expected value of X_R, is more general than has been presented in the past. Both Widrow and Lowell considered a rounding lattice with $c = 0$. Also using (2.2–9) and (2.2–10) we can obtain the $V(X_R)$, which will be required in Chapter 6.

$$V[X_R] = \sigma^2 + \frac{w^2}{12} + 4\sum_{k=1}^{\infty}(-1)^k\left[\sigma^2+\left[\frac{w}{2\pi k}\right]^2\right]\exp\left[-2\left[\frac{\pi k\sigma}{w}\right]^2\right]\mathrm{Cos}\left[2\pi k\left[\frac{\mu}{w} - c\right]\right]$$

$$- \left[\frac{w}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\exp\left[-2\left[\frac{\pi k\sigma}{w}\right]^2\right]\mathrm{Sin}\left[2\pi k\left[\frac{\mu}{w} - c\right]\right]\right] \quad (2.2\text{-}15)$$

For $c = 0$, expressions (2.2–13) and (2.2–15) agree with the mean and variance of X_R given in Lowell (1980).

In considering the effect of rounding on the normal distribution it will be assumed, without loss of generality that the mean is zero. The expressions for the moments of X_R will be simplified and α becomes equal to $-c$. For a normal distribution with mean zero and variance σ^2, from (2.2-9) to (2.2-12) the first four moments of X_R are:

$$\left.\begin{aligned}
E[X_R] &= -\sigma\left[\frac{r}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\,D\,\mathrm{Sin}(2\pi kc)\right]\Bigg] \\
E[X_R^2] &= \sigma^2\left[1+\frac{r^2}{12}+4\sum_{k=1}^{\infty}(-1)^k\left[1+\left[\frac{r}{2\pi k}\right]^2\right]D\,\mathrm{Cos}(2\pi kc)\right] \\
E[X_R^3] &= \sigma^3\left[\frac{3r}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\left[1+\left[\frac{2\pi k}{r}\right]^2+\frac{1}{2}\left[\frac{r}{\pi k}\right]^2-\frac{r^2}{12}\right]D\,\mathrm{Sin}(2\pi kc)\right. \\
E[X_R^4] &= \sigma^4\left[3+\frac{r^2}{2}+\frac{r^4}{80}+\frac{r}{\pi}\sum_{k=1}^{\infty}\left[\left[\frac{r}{\pi k}\right]\left[\frac{r^2}{2}-6\right]^2-3\left[\frac{r}{\pi k}\right]^3-32\left[\frac{\pi k}{r}\right]^3\right.\right. \\
&\qquad \left.\left.+2r\pi k\right]D\,\mathrm{Cos}(2\pi kc)\right.
\end{aligned}\right\}(2.2\text{-}16)$$

where $D = \exp\left[-\frac{2\pi^2k^2}{r^2}\right]$

A Fortran program was written to calculate the first four moments (2.2-16) and related measures σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R}. The rounding precision r varied upto 5, for lattice positions c = -0.5, -0.45,..., 0.45, 0.5. The ranges of the first four moments of X_R are given in Table (2.2.1). These are compared with results (2.2-5) given by Sheppard's corrections. Table (2.2.2) shows the maximum bias caused by rounding in the measures μ, σ^2, $\sqrt{\beta_1}$ and β_2. The basis for each parameter is defined as:

$$B_1 = \frac{\mu_R - \mu}{\sigma} \qquad B_2 = \frac{\sigma_R^2 - \sigma^2}{\sigma^2}$$

$$B_3 = \sqrt{\beta_{1R}} - \sqrt{\beta_1} \qquad B_4 = \beta_{2R} = \beta_2 \qquad (2.2\text{-}17)$$

The maximum bias is simply the maximum value of B_i.

The range in the values of the moments is caused by the lattice position c. As the moments of X_R were obtained for 21 values of c, the results in Table (2.2.1) and Table (2.2.2) are an indication of the possible range of the moments and maximum bias respectively. The values for μ_R and σ^2_R in Table (2.2.2) agree with Widrow (1961) for $r \leq 1.0$. For $r > 1.0$ they differ, as Widrow uses an approximation to find μ_R and σ^2_R and considers only lattice position $c = 0$. The results in Table (2.2.2) demonstrate how the lattice effect increases considerably as the rounding becomes more coarse. This being more noticeable in the 2nd and 4th moments. Sheppard's corrections provide a very good approximation to the moments and related measures of X_R for $r \leq 1.0$. For $r > 1$ the lattice effect is seen to cause these corrections to become less effective in adjusting for rounding and are unreliable for $r > 2.0$.

As r increases in size the lattice position (or position of the mean μ relative to nearest midpoint) is crucial in deciding how much the normal distribution is distorted by rounding. This can be illustrated by examining the biases in the four measures μ, σ^2, $\sqrt{\beta_{1R}}$ and β_{2R} due to rounding over the range of c. Figures (2.2.1) to (2.2.4) show the curves for these baises for c ranging between $-\frac{1}{2}$ and $\frac{1}{2}$ and r up to 5.0 (except for B_4 when r is upto 4). These Figures illustrate how the biases are dependent on the value of c. In the graph for B_1, the bias is zero for $c = 0$ and $\pm \frac{1}{2}$. In general, whenever the mean coincides with the boundary

or centre of a rounding interval then the bias is zero. For the variance, the bias B_2 is symmetrical about c = 0. When c is between ± 0.2 the value of σ^2_R can be less than that of σ^2. The graph for B_3 is very similar to that for B_1, again having zero bias at c = 0 and ± ½. The interesting feature about the bias caused by rounding in β_2 is that, as r rises in value, B_4 rapidly increases as c approaches zero. This is caused by the fact that σ^2_R tends to zero in the region of c = 0 for large r.

In Tricker (1984) the bias in the mean and variance is given with respect to the lattice position for a Laplace distribution which has been subject to rounding. This bias in the mean was found to be generally larger than for the normal, where the bias variance was similar.

Table 2.2.1

Range of the first 4 moments of rounded normal data (Sheppard's Corrections in brackets)

r	$E[X_R]/\sigma$	$E[X_R^2]/\sigma^2$
1	$\pm 8.10(10)^{-10}$, (0.0)	1.0833-1.0833(1), (1.0833)
2	$\pm 4.36(10)^{-3}$, (0.0)	1.3077-1.3650, (1.3333)
3	$\pm 1.01(10)^{-1}$, (0.0)	1.2020-2.2986, (1.7500)
4	$\pm 3.53(10)^{-1}$, (0.0)	0.6964-4.0020, (2.3333)
r	$E[X_R^3]/\sigma^3$	$E[X_R^4]/\sigma^4$
1	$\pm 9.80(10)^{-8}$, (0.0)	3.5125-3.5125(1), (3.5125)
2	$\pm 1.40(10)^{-1}$, (0.0)	4.6745-5.7255, (5.200)
3	± 1.55, (0.0)	6.1559-10.8500, (8.5125)
4	± 3.19, (0.0)	11.9813-16.0811, (14.2000)

(1) no range in values correct to four decimal places.

Table 2.2.2

Maximum bias in μ, σ^2, $\sqrt{\beta_1}$ and β_2 caused by rounding (Sheppard's Corrections in brackets) for normal distribution

r	B_1	B_2	B_3	B_4
1	$8.10(10)^{-10}$ (0.0)	$8.33(10)^{-2}$ $(8.33(10)^{-2}\sigma^2)$	$8.90(10)^{-8}$ (0.0)	$7.10(10)^{-3}$ $(7.10(10)^{-3}$
2	$4.36(10)^{-3}$ (0.0)	$3.67(10)^{-1}$ $(3.33(10)^{-1}\sigma^2)$	$1.01(10)^{-1}$ (0.0)	$4.91(10)^{-1}$ $(7.5(10)^{-2})$
3	$1.01(10)^{-1}$ (0.0)	1.30 $(7.5(10)^{-1}\sigma^2)$	1.03 (0.0)	4.51 $(2.2(10)^{-1})$
4	$3.73(10)^{-1}$ (0.0)	3.00 $(1.33\sigma^2)$	2.22 (0.0)	21.10 $(3.9(10)^{-1})$

Figure 2.2.1

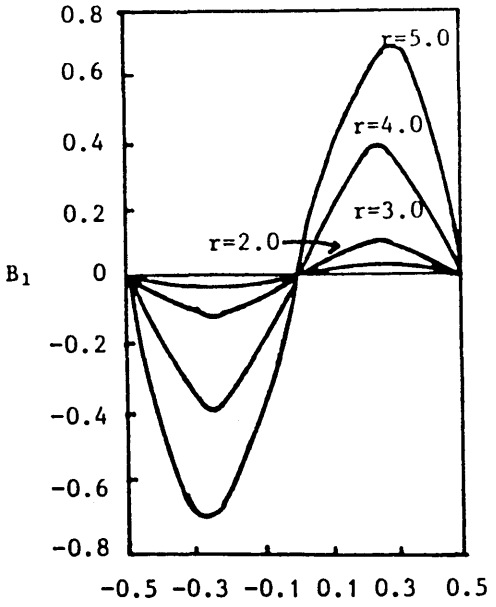
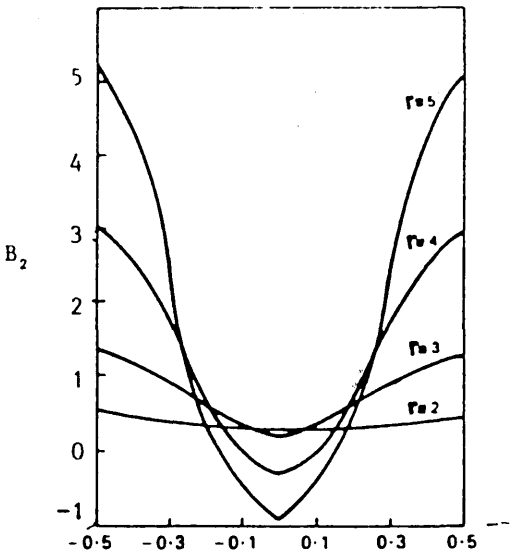
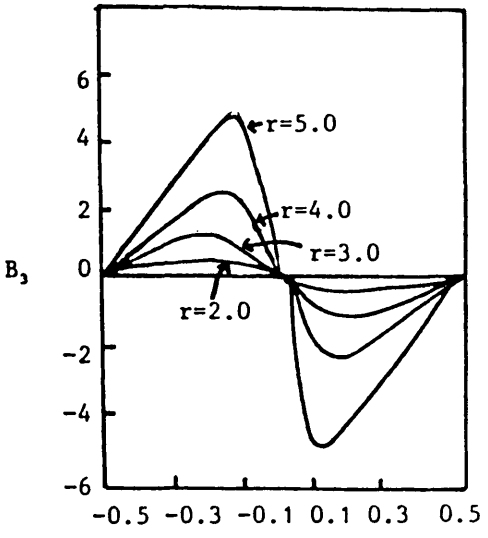


Figure 2.2.2





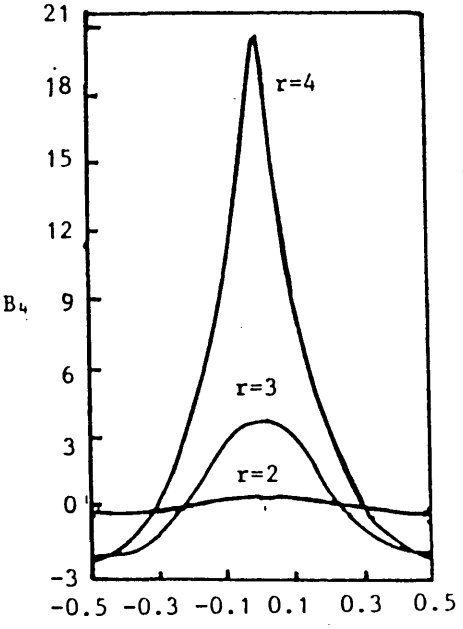


Figure 2.2.3

Figure 2.2.4

Figure 2.2.1 B_1 for values of c between $\pm \frac{1}{2}$ and r ranging upto 5

Figure 2.2.2 B_2 for values of c between $\pm \frac{1}{2}$ and r ranging upto 5

Figure 2.2.3 B_3 for values of c between $\pm \frac{1}{2}$ and r ranging upto 5

Figure 2.2.4 B_4 for values of c between $\pm \frac{1}{2}$ and r ranging upto 4

Gamma Distribution

Past research into the moments of rounded non-normal distributions has concentrated on finding specific corrections to moments where Sheppard's corrections are known to break down. Most of this work was carried out before the Second World War by such authors as Pairman & Pearson (1918), Pearse (1928), Martin (1934) and Sandon (1924). These corrections are very difficult to make. This section shows how the general result for the moments of X_R given in section (2.2.1) can be applied to a non-normal distribution. Secondly it shows the importance of the shape of the distribution in determining the effect of rounding. No attention has previously been given to this. Only Holland (1975) when investigating the bias in the mean and variance of rounded data points out that skewness of a distribution may influence the rounding effect. It shall be demonstrated by using the gamma distribution how the rounding bias in μ, σ^2, $\sqrt{B_1}$ and β_2 is strongly dependent on the shape of the distribution.

Let the random variable X have a gamma distribution with the following p.d.f.

$$f(x) = \frac{1}{\Gamma(\alpha)} \frac{1}{\theta} \left[\frac{x}{\theta}\right]^{\alpha-1} e^{-\frac{x}{\theta}} \qquad \begin{matrix} x \geq 0 \\ \theta,\ \alpha > 0 \end{matrix} \qquad (2.2\text{-}18)$$

with mean $\alpha\theta$ and variance $\sigma^2 = \alpha\theta^2$.

Let X_R be the random variable corresponding to a rounding lattice with interval of width w and lattice position c. Using the CF of (2.2-18) we obtain from (2.2-9) to (2.2-12) expressions for the first four moments of X_R

$$E[X_R] = E[X] + \sigma\left[\frac{r}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}G^{-\frac{\alpha}{2}}\sin(\alpha\psi-2\pi kc)\right]$$

$$E[X_R^2] = E[X^2] + \sigma^2\left[\frac{r^2}{12} + \left[\frac{r}{\pi}\right]^2\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}G^{-\frac{(\alpha+1)}{2}}\right. \qquad (2.2\text{-}19)$$

$$\left.\times\left[2\left[\frac{\pi k}{r}\right]\sqrt{\alpha}\,\sin[(\alpha+1)\psi-2\pi kc] + G^{\frac{1}{2}}\cos(\alpha\psi-2\pi kc)\right]\right]$$

$$E[X_R^3] = E[X^3] + \sigma^3\left[\frac{r\sqrt{\alpha}}{4} - 3\left[\frac{r}{\pi}\right]^3\sum_{k=1}^{\infty}\frac{(-1)^k}{k^3}G^{-\frac{(\alpha+2)}{2}}\right.$$

$$\left[(\alpha+1)\left[\frac{\pi k}{r}\right]^2\sin[(\alpha+2)\psi-2\pi kc] - \left[\frac{\pi k}{r}\right](\alpha G)^{\frac{1}{2}}\cos[(\alpha+1)\psi-2\pi kc]\right.$$

$$\left.\left.+\frac{1}{12}\left[6-(\pi k)^2\right]G\,\sin(\alpha\psi-2\pi kc)\right]\right]$$

$$E[X_R^4] = E[X^4] = \sigma^4\left[\frac{r^2(\alpha+1)}{2} + \frac{r^4}{80} + \left[\frac{r}{\pi}\right]^4\sum_{k=1}^{\infty}\frac{(-1)^k}{k^4}G^{-\frac{(\alpha+3)}{2}}\times\right.$$

$$\left[4\left[\frac{\pi k}{r}\right]^3\frac{(\alpha+1)(\alpha+2)}{r\alpha}\sin[(\alpha+3)\psi-2\pi kc]\right.$$

$$+ 6\left[\frac{\pi k}{r}\right]^2(\alpha+1)G^{\frac{1}{2}}\cos[(\alpha+2)\psi-2\pi kc]$$

$$- \left[\frac{\pi k}{r}\right][6-(\pi k)^2]\sqrt{\alpha}G\sin[(\alpha+1)\psi-2\pi kc]$$

$$\left.\left.+\frac{1}{2}\left[(\pi k)^2-6\right]G^{3/2}\cos(\alpha\psi-2\pi kc)\right]\right]$$

where $G = \left[1 + \frac{(2\pi k)^2}{\alpha r^2}\right]$, $\tan\psi = \frac{2\pi k}{r\sqrt{\alpha}}$.

The shape of the distribution will be an important determinant of the effect of rounding on the moments. For example the dominant term in $E[X^2_R]$ is $G^{-(\alpha+1)/2}$, where α is the shape parameter of the gamma distribution. For fixed r, $G^{-(\alpha+1)/2}$ will approach $E[X^2] + \sigma^2 r^2/12$ (Sheppard's correction) as the distribution becomes more symmetrical. In general we would expect the moments of X_R to tend towards Sheppard's corrections as the distribution becomes less skewed.

A Fortran program was written to calculate the first four moments (2.2-19) and related measures σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R}. The rounding precision r varied up to 4, for lattice positions c = -0.5, -0.45, ..., 0.45, 0.5. For the gamma distribution, the rounding lattice where c is less than zero, may cause data to be rounded to values outside of its range. This is really a theoretical point, because in practice rounding lattice with this value of c would be uncommon. Generally for distributions which have a finite terminal at the end of one of its ranges, the rounding lattice may cause data to be rounded to values outside its range. Selected results are given in Tables (2.2.3) and (2.2.4). Only results for the related measures are shown, as their general behaviour with respect to the shape of the distribution will be similar to that of the moments.

The results in Table (2.2.3) indicate how crucial the shape of the distribution is in determining the rounding effect. As the distribution becomes more skewed the lattice effect increases, causing a greater bias in the rounded parameters μ_R, σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R}. As implied by the expressions for the moments of X_R, as α increase, the values of these parameters tend towards those given by Sheppard's corrections. For $\alpha \geq 5$ Sheppard's corrections are reliable for $r \leq 1.0$. Table (2.2.4) shows the maximum bias that may occur in the parameters. For the

exponential distribution (α = 1) this bias may be severe, but quickly reduces as α increases. As expected for increasing α, the bias in the parameters approaches those of the normal distribution. Figures (2.2.5) and (2.2.6) show curves for B_1 and B_2 for c ranging between $-\frac{1}{2}$ and $\frac{1}{2}$, r = 1 and α up to 4. These illustrate how B_1 and B_2 can vary considerably for different lattice positions. The same was also found true for B_3 and B_4.

Table 2.2.3

Range in mean, variance, skewness and kurtosis for gamma rounded and unrounded data. (Sheppard's corrections in brackets).

α	r	μ_R	σ^2_R	$\sqrt{\beta_{1R}}$	β_{2R}
	0	1.00	1.00	2.00	9.00
1	0.5	0.99-1.02 (1.0)	0.98-1.04 (1.02)	1.88-2.06 (1.94)	8.52-9.24 (8.76)
	1.0	0.96-1.08 (1.0)	0.92-1.17 (1.08)	1.55-2.24 (1.77)	7.30-10.05 (8.11)
	2.0	0.84-1.31 (1.0)	0.73-1.72 (1.33)	0.67-3.01 (1.30)	4.40-14.3 (6.30)
	0	2.23	1.00	0.89	4.20
5	0.5	2.24-2.24 (2.23)	1.02-1.02 (1.02)	0.87-0.87 (0.86)	4.20-4.20 (4.15)
	1.0	2.23-2.24 (2.23)	1.08-1.09 (1.08)	0.77-0.81 (0.79)	4.00-4.03 (4.02)
	2.0	2.19-2.28 (2.23)	1.16-1.50 (1.33)	0.25-1.04 (0.58)	2.82-3.98 (3.60)
	0	4.47	1.00	0.45	3.30
20	0.5	4.47-4.47 (4.47)	1.02-1.02 (1.02)	0.44-0.44 (0.44)	3.30-3.30 (3.29)
	1.0	4.47-4.47 (4.47)	1.08-1.08 (1.08)	0.40-0.40 (0.40)	3.20-3.30 (3.25)
	2.0	4.48-4.48 (4.47)	1.27-1.40 (1.33)	0.11-0.48 (0.29)	2.31-3.20 (3.09)

Note: All values in table for μ_R and σ^2_R are multiples of σ and σ^2 respectively and correct to two decimal places.

Table 2.2.4

Maximum bias expected for mean, variance, skewness and kurtosis caused by rounding in gamma distribution.

	B_1		B_2		B_3		B_4	
α \ r	1.0	2.0	1.0	2.0	1.0	2.0	1.0	2.0
1.0	$8.2(10)^{-2}$	$3.1(10)^{-1}$	17.5	72.0	$4.4(10)^{-1}$	1.3	1.7	5.3
5.0	$0.1(10)^{-2}$	$4.2(10)^{-2}$	8.9	49.6	$1.1(10)^{-1}$	$6.4(10)^{-1}$	$4.4(10)^{-1}$	1.4
10	$1.1(10)^{-6}$	$2.1(10)^{-2}$	8.4	43.1	$7.4(10)^{-2}$	$4.7(10)^{-1}$	$1.1(10)^{-1}$	1.2
30	$1.0(10)^{-6}$	$8.7(10)^{-3}$	8.3	38.5	$4.1(10)^{-2}$	$2.8(10)^{-1}$	$4.2(10)^{-2}$	$(8.1)(10)^{-1}$
40	$5.2(10)^{-9}$	$7.7(10)^{-3}$	8.3	38.1	$3.6(10)^{-2}$	$2.6(10)^{-1}$	$3.4(10)^{-2}$	$7.2(10)^{-1}$
Normal	$8.1(10)^{-10}$	$4.4(10)^{-3}$	8.3	36.7	$7.6(10)^{-8}$	$1.1(10)^{-1}$	$7.1(10)^{-3}$	$4.9(10)^{-1}$

For the two distributions considered in this section the CFs had a closed form and the values of A^S and B^S in the expressions for the moments of X_R presented no problems. Not all distributions have a CF with a closed form and the values of A^S and B^S will have to be obtained by numerical integration or series expansion. With today's computing facilities this should not present much of a problem.

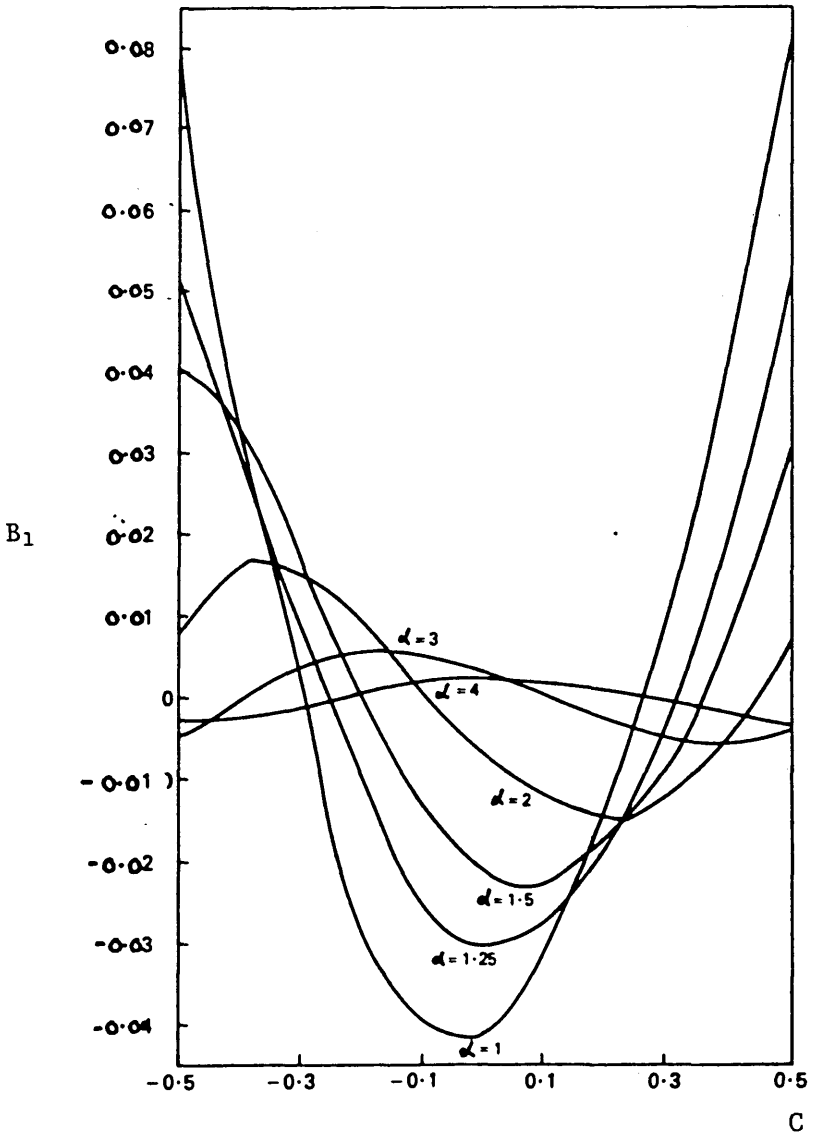


Figure 2.2.5 B_1 for values of c between $\pm \frac{1}{2}$, r = 1, and α upto 4 for gamma distribution

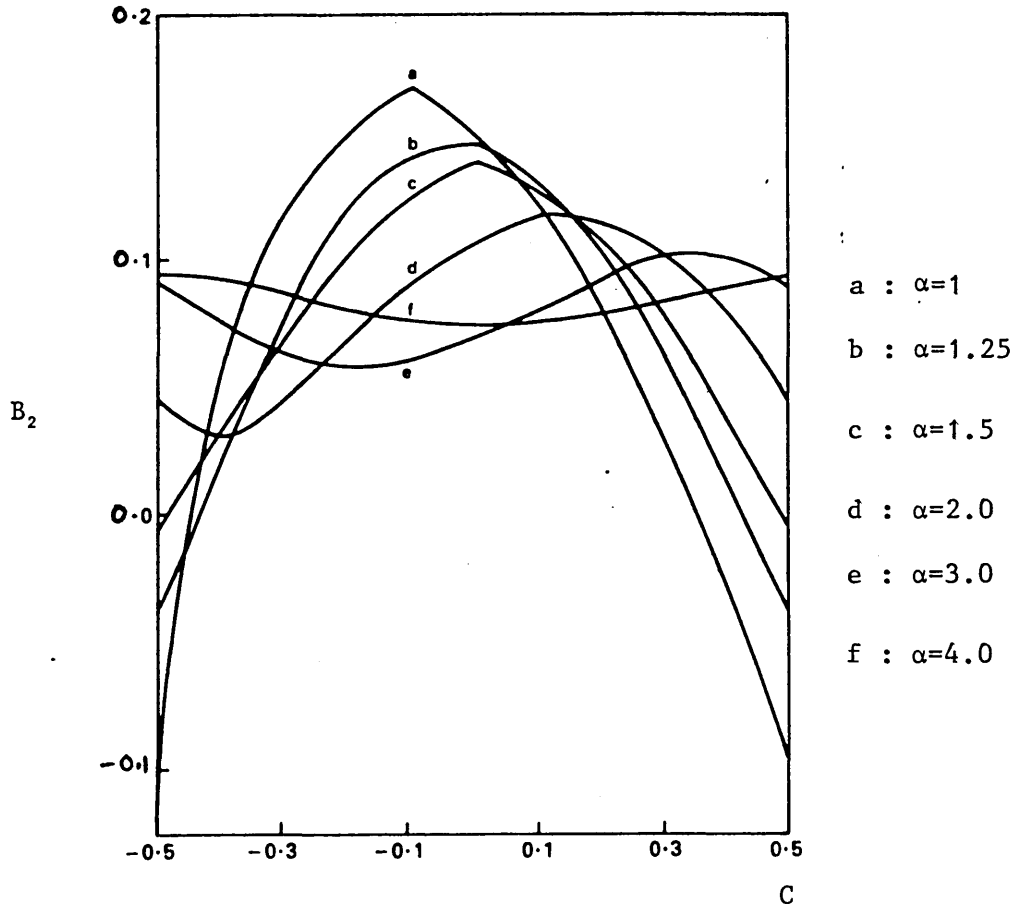


Figure 2.2.6 B_2 for values of c between ± ½, r = 1 and α upto 4 for gamma distribution

2.2.3 Relationship between the shape of a distribution and the effect of rounding on its moments

The two distributions studied in the previous section indicate that the moments of the normal distribution are very robust to the process of rounding, while departure from non-normality can cause rounding to have an increased effect. In practice we often deal with non-normal distributions and hence the effect of rounding on the moments of such distributions is of interest. An example is when dealing with the effect of non-normality on test statistics. The influence of non-normality on these tests is to some extent determined by the values of $\sqrt{\beta_1}$ and β_2 in the population. The degree to which rounding may change these two parameters under non-normality is important, as this will influence the behaviour of the test statistic.

As mentioned in the previous section, there has been research into the effect of rounding on the moments of a normal distribution, but none concerned with the effect of non-normality. For the first time the association between the bias caused by rounding in the moments and the shape of the distribution will be considered. This present section aims to set out in an unsophisticated way the relationship between the change in moments due to rounding and the shape of the distribution. This relationship will be presented diagrammatically to make it easy to assimilate. Results will be given for the moments μ and σ^2 and moment ratios $\sqrt{\beta_1}$ and β_2. These have been chosen as they are usually of more practical use than the first four moments about zero.

To study the association between rounding on the moments and shape of distribution, it is helpful to look at a system of distributions. The best known is the Pearson System, which covers many well known statistical distributions.

However the Pearson family lacks a clear systematic basis and would be more difficult for examining the effect of rounding. Another family of distributions that could be used is the Bessel Function Distributions introduced by McKay (1932). The advantage of this system is that there is a standard closed expression for the CF and thus the moments of X_R can be easily obtained from (2.2-9) to (2.2-12). However the system is restricted in the (β_1,β_2) plane. This system of distributions will provide one distribution for $(\beta_2-3)\beta_1 > 1.5$ ie below the type III line. A system of distributions that is convenient to use and covers the (β_1,β_2) plane is the Johnson System. Johnson (1949) described a system that permits a simple transformation of a normal variate such that, for any possible pair of values β_1,β_2 there is just one member of this family. In fact there are three transformations, which are defined in Table (2.2.5).

Table (2.2.5)

Transformations for Johnson Distributions, where z denotes a standard normal variate

Name	Form of variate x	Range
Johnson S_B	$z = \gamma + \delta\log[x/(1-x)]$	(0,1)
Johnson S_U	$z = \gamma + \delta\sinh^{-1}x$	$(-\infty,\infty)$
Lognormal S_L	$z = \gamma + \delta\log x$	$(0,\infty)$

The Johnson System was used for two main reasons. It was simple to use and covered the $(\beta_1\beta_2)$ plane. Secondly it would be useful to use the computer tracings of the various Johnson distributions given by Pearson and Please (1975). An advantage in using the Pearson System would be that it covers well known

statistical distributions. However authors such as Pearson and Please (1975) have demonstrated that the Johnson System can be used in place of the Pearson System with very little difference in results. The relationship between the Pearson and Johnson Systems is given in Figure (2.2.7) for the region of the (β_1,β_2) considered in this study.

It was decided to use the same set of 29 Johnson distributions as used by Pearson and Please (1975). Four more were added to this set to include distributions that were U shaped. These being $\beta_2 = 2.0$, $\sqrt{\beta_1} = 0.8$ and $\beta_2 = 1.7$, where $\sqrt{\beta_1} = 0.0, 0.2, 0.4$. A computer program was written to obtain an outline of these four distributions. This set of 33 distributions adequately covers the various distribution shapes that may be seen in practice. Each of the distributions was standardised to have a mean zero and variance one. This simplifies the rounding, as r = w and the distance from the mean to the nearest midpoint is the lattice position c. We lose no generality in the results by standardizing the distributions. The curves for the standardized Johnson distributions are shown in Figures (2.2.8) and (2.2.9). The dashed curves are normal distributions with mean zero and standard deviation one.

The main advantage in using distributions from Johnson's System lies in the simple relationship between their variables and a standard normal variable. This means that in terms of simulation purposes they are very efficient to use. For this reason, the behaviour of the moments for rounded data was investigated by simulation.

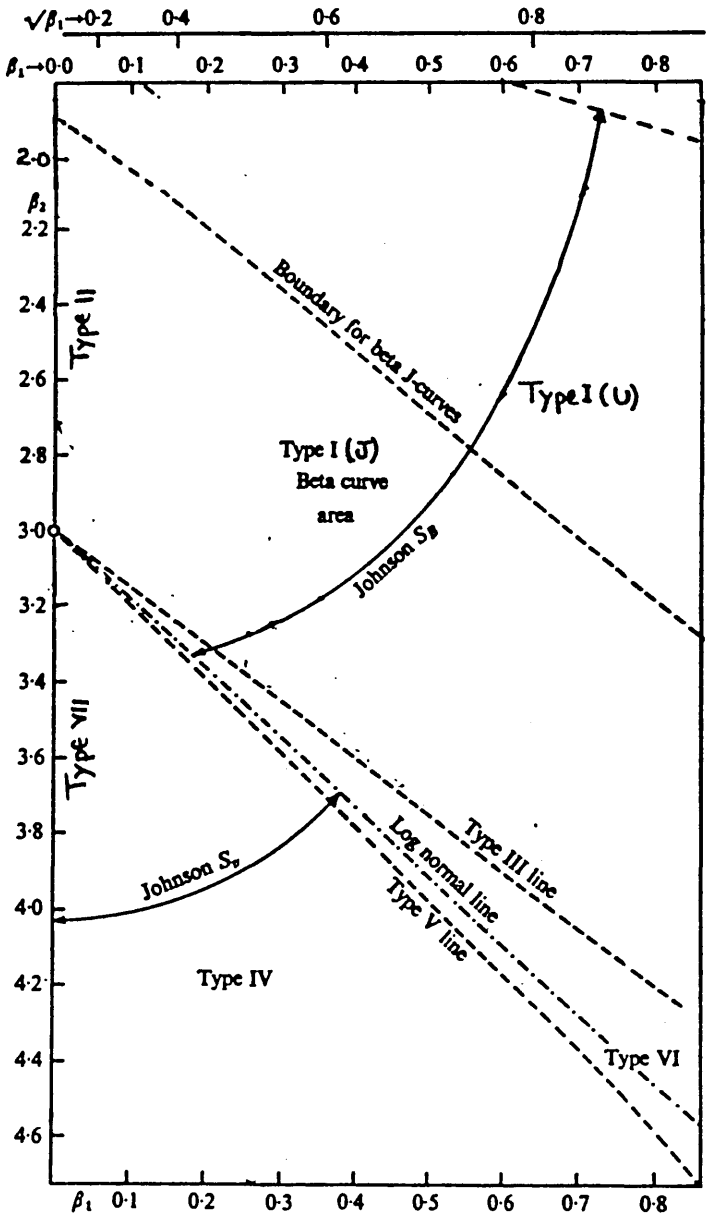


Figure 2.2.7 Regions and boundaries in $(\beta_1 \beta_2)$ plane of Pearson and Johnson distributions

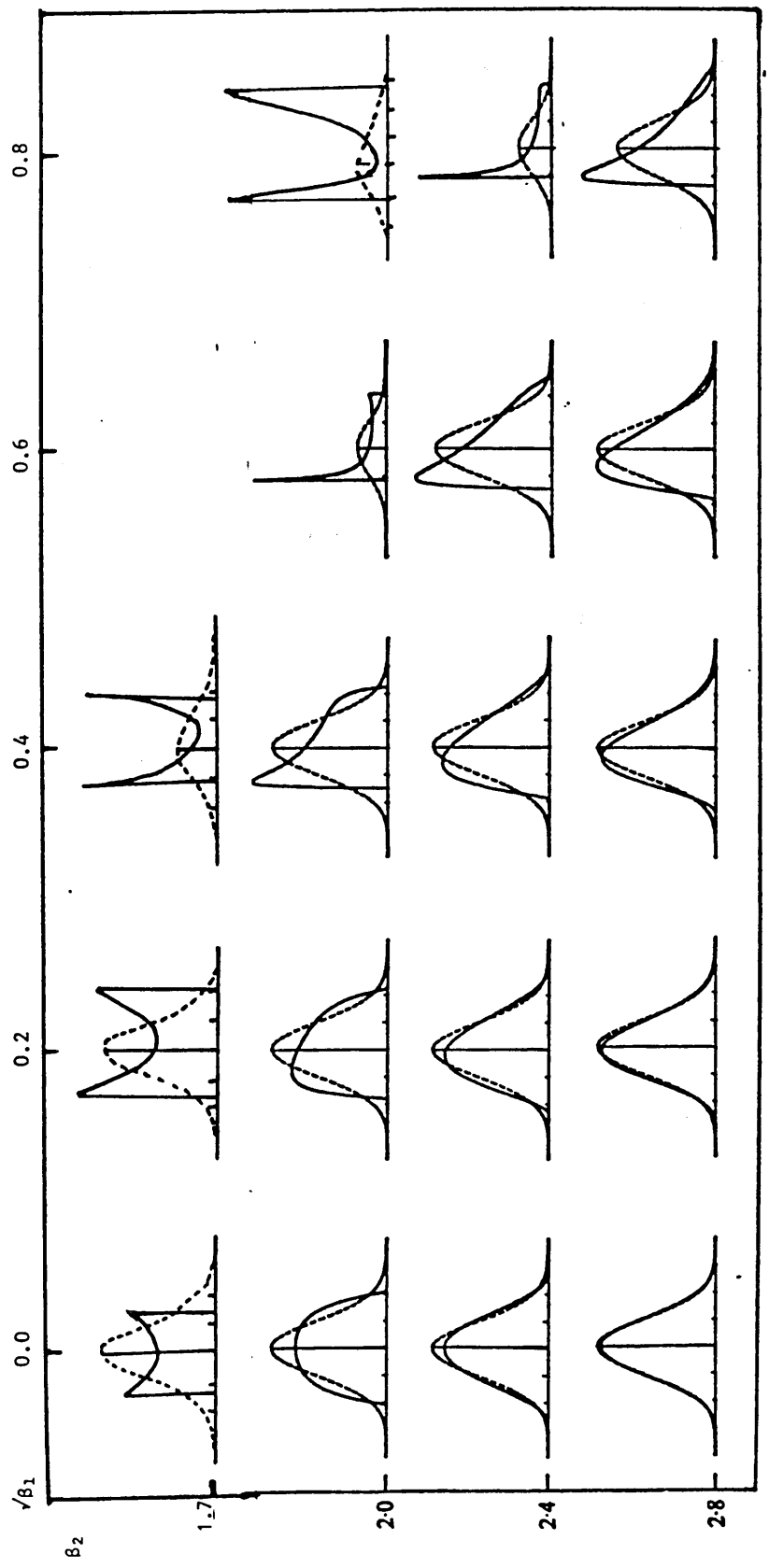


Figure 2.2.8 Curve for Johnson distributions – the dashed curves are standard normal distributions

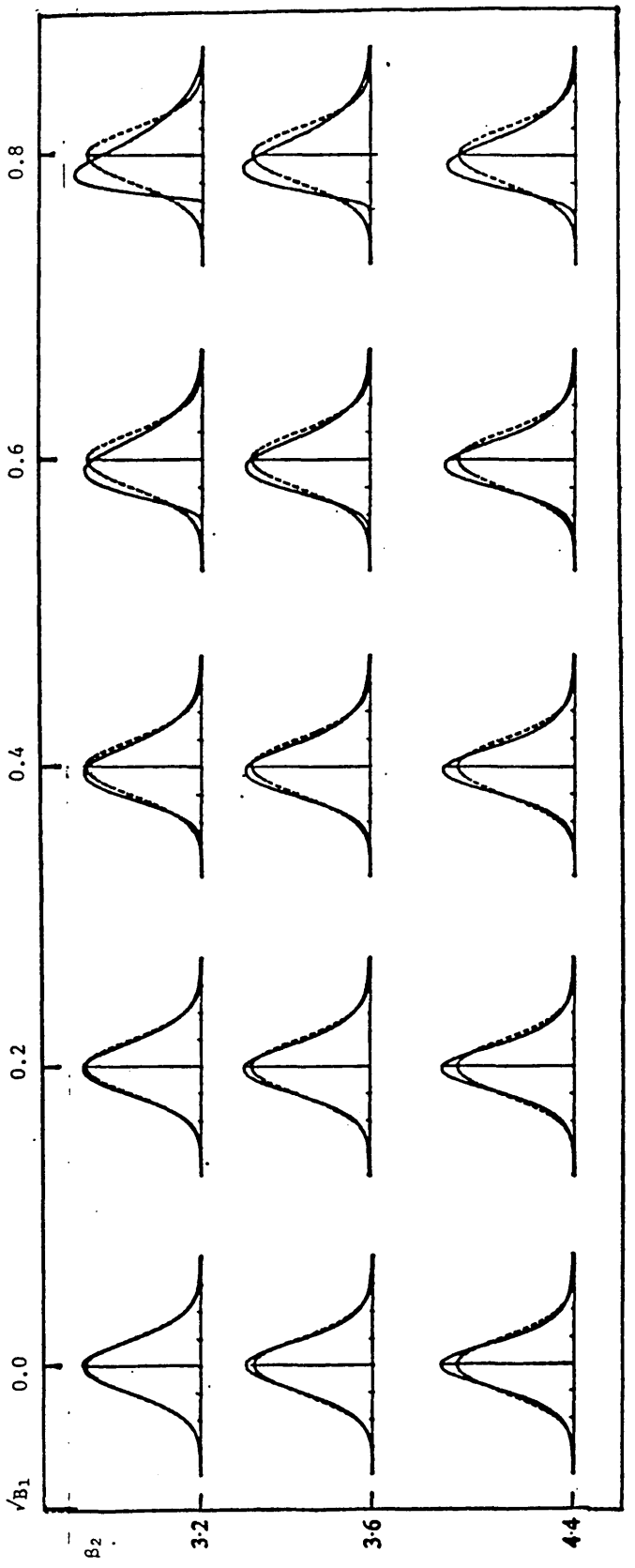


Figure 2.2.9 Curves for Johnson distributions, the dashed curves are standard normal distributions

Simulation Study

A Fortran program JMOMENTS was written for the simulation. Deviates from each of the 33 standardized Johnson distributions were obtained using the transformations given in Table (2.2.5) with NAG routines used to generate the standard normal deviates. The Johnson deviates were rounded according to a rounding lattice with rounding interval w and lattice position c. For various combinations of r and c each Johnson deviate was obtained 100,000 times. From these 100,000 replicates the mean, variance, skewness and kurtosis of the rounded Johnson distribution were obtained. The rounding precision varied upto 2, for lattice positions c between $\pm\frac{1}{2}$.

In order to simulate Johnson deviates which had a distribution with specific β_1, β_2, we required the values of the parameters δ, γ in Table (2.2.5). Unfortunately Pearson and Please (1975) did not provide these. The parameters δ and γ were obtained by using tables from Pearson and Hartley (1972) Vol 2.

The program was fully tested. For example, to check the generation of Johnson deviates, the values of μ, σ, $\sqrt{\beta_1}$ and β_2 given by the simulation were compared with their expected values. Appendix A gives a list of all the output produced by the JMOMENTS program.

Discussion of Simulation Results

A major problem in this study was to summarise the large number of results. These consisted of the parameters μ_R, σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R} for each of the 33 distributions, each of 3 values of r and 11 lattice positions. The results are given

in the form of contour diagrams so that they can be more easily understood. Each contour diagram shows the maximum bias expected in the parameter, over the region $0 \leq \sqrt{\beta_1} \leq 0.8$ and $1.7 < \beta_2 < 4.4$ for a given r. The bias for each parameter is defined as (2.2–16).

Before considering the results a point should be made concerning the outlines of the Johnson distributions in Figures (2.2.8) and (2.2.9). For $\beta_2 < 3$ a value of $\sqrt{\beta_1}$ will cause a greater degree of non–normality than for the same value of $\sqrt{\beta_1}$ when $\beta_2 > 3$. Clearly as shown by the outlines, it is the joint values of $\sqrt{\beta_1}$ and β_2 that determine the degree of non–normality.

The results are presented in figures (2.2–10) to (2.2–12). They provide a good evaluation of the bias expected in the parameters μ_R, σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R} over a practical range of (β_1,β_2). The figures speak for themselves and the reader may readily determine the general trend of what happens to the bias for different values of (β_1,β_2). However a number of obvious remarks can be made concerning the results given in the figures. A striking feature is the extent to which the departure from normality determines the bias caused by rounding. The most important feature of non–normality being the extent to which the distribution departs from symmetry. The largest biases are found in the top right hand corner of the $(\beta_1\beta_2)$ region considered, where the departure from symmetry is greatest.

Sheppard's corrections are reliable in the region of (β_1,β_2), where the departure from normality is less. The size of this region is determined by the value of r. For the mean, this region is about 80% of (β_1,β_2) region under consideration for r = 0.5 and only 17% for 1.5, whereas for the variance, Sheppard's corrections are reliable for about 80% of (β_1,β_2) region under consideration for r = 0.5 and

falls to 35% for r = 1.5. It has been customary to assume that Sheppard's corrections are reliable if the distribution has high order contact. The Johnson System of distributions all have high order contact and thus would expect Sheppard's corrections to be reasonably reliable. However, as the results indicate this is not always true.

Figure 2.2.10 Contour diagrams for the maximum bias in mean (B_1) for $\sqrt{\beta_1}$ and β_2 [bias given by Sheppard's correction is zero]

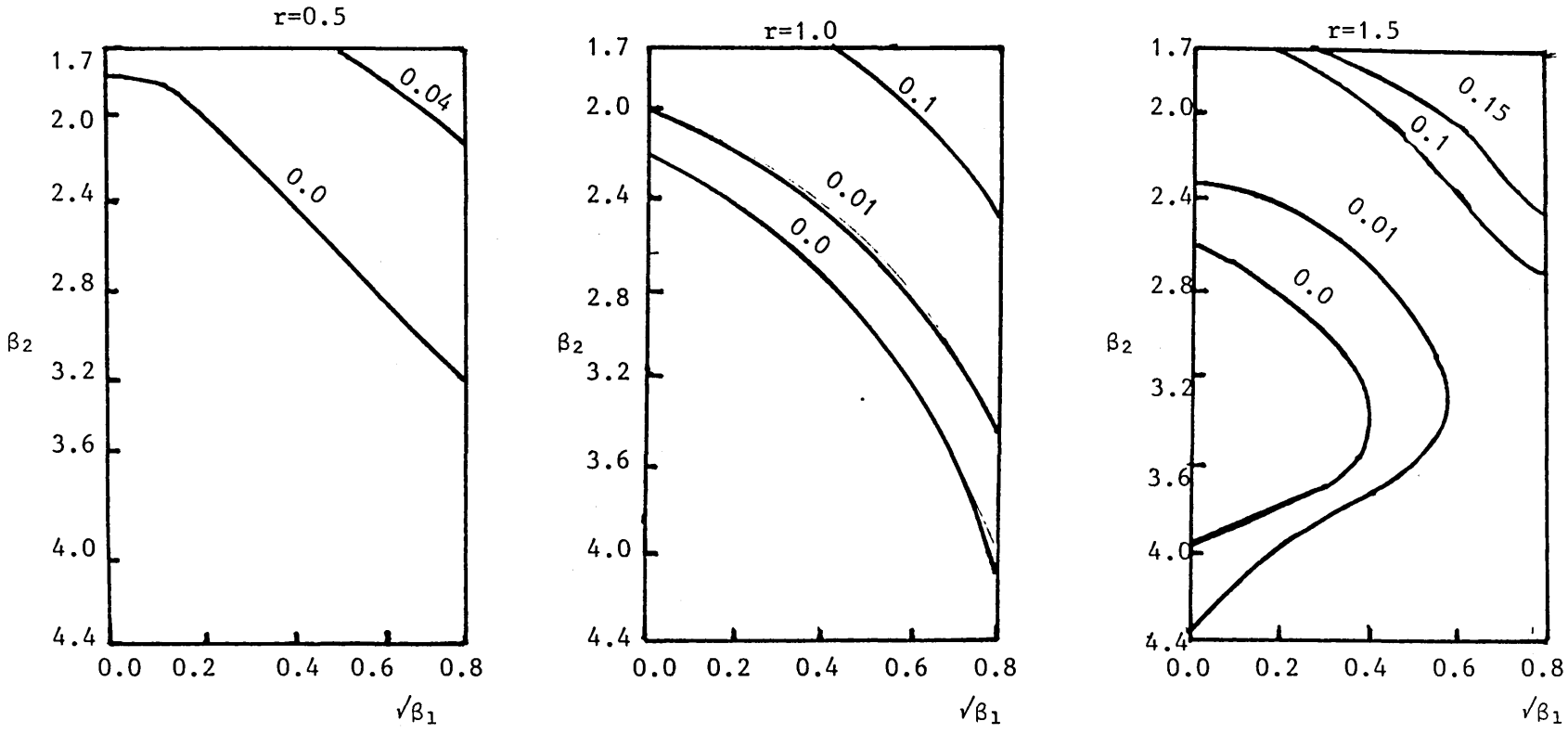


Figure 2.2.11 Contour diagrams for the maximum bias in the variance (B_2) for $\sqrt{\beta_1}$ and β_2

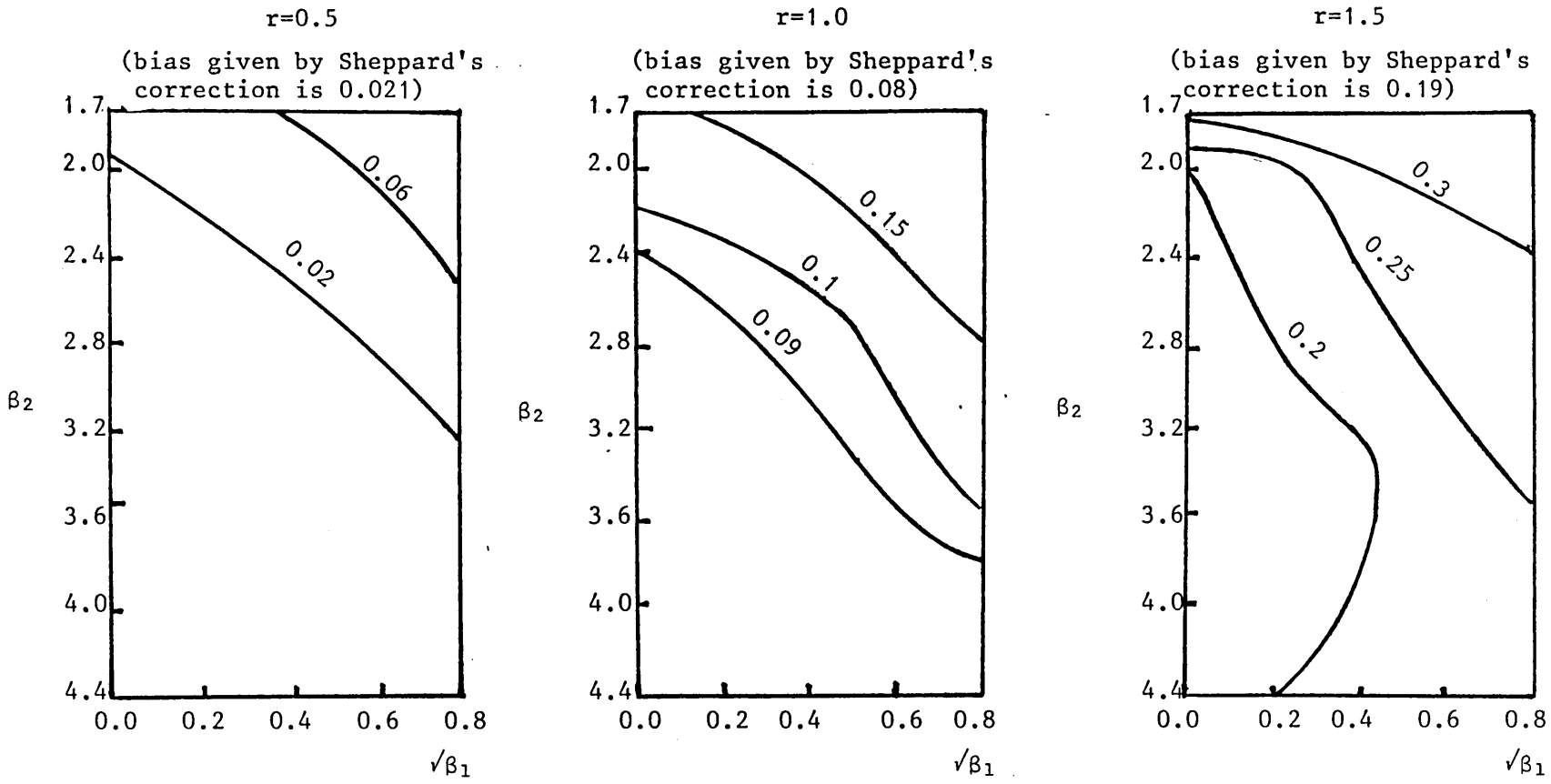


Figure 2.2.12 Contour diagrams for the maximum bias in $\sqrt{\beta_1}$ (B_3) for $\sqrt{\beta_1}$ and β_2

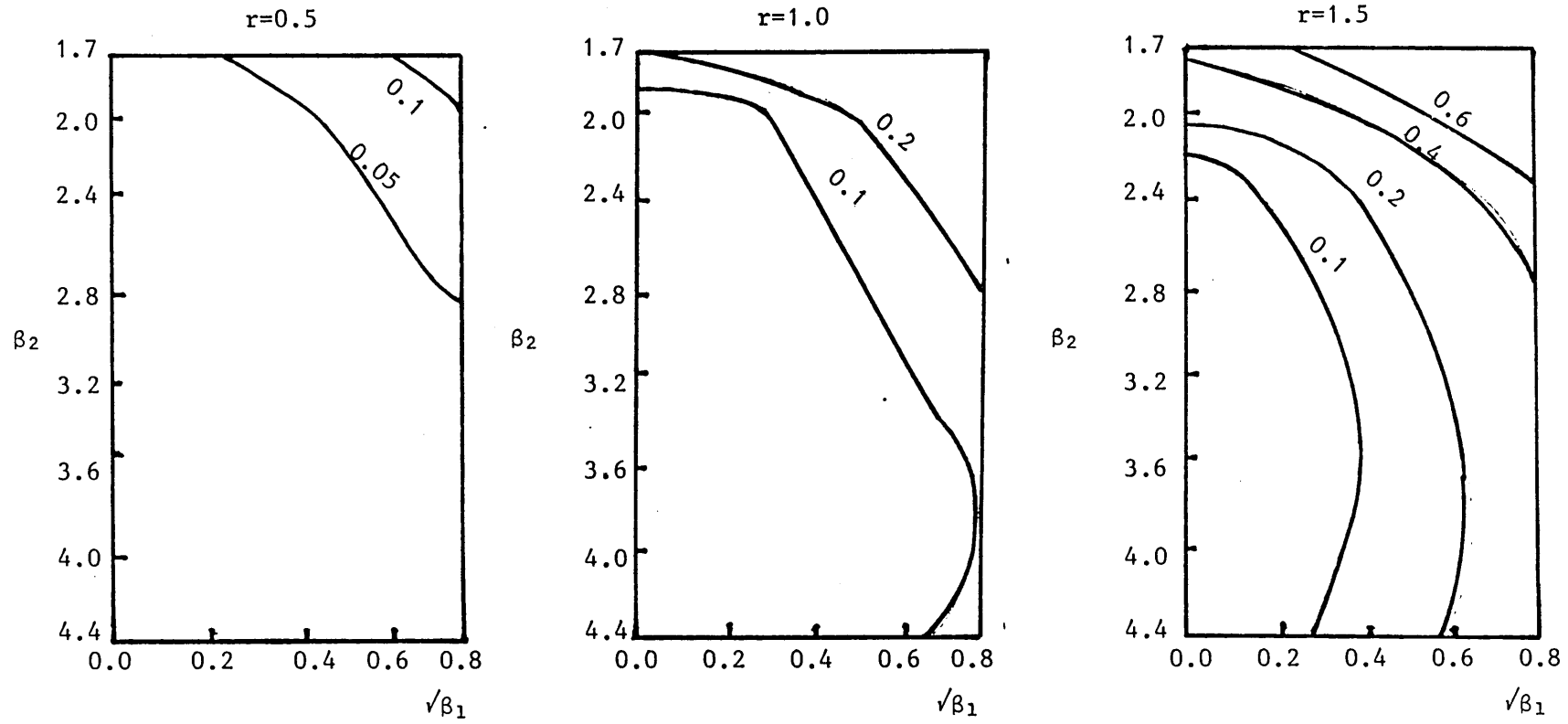
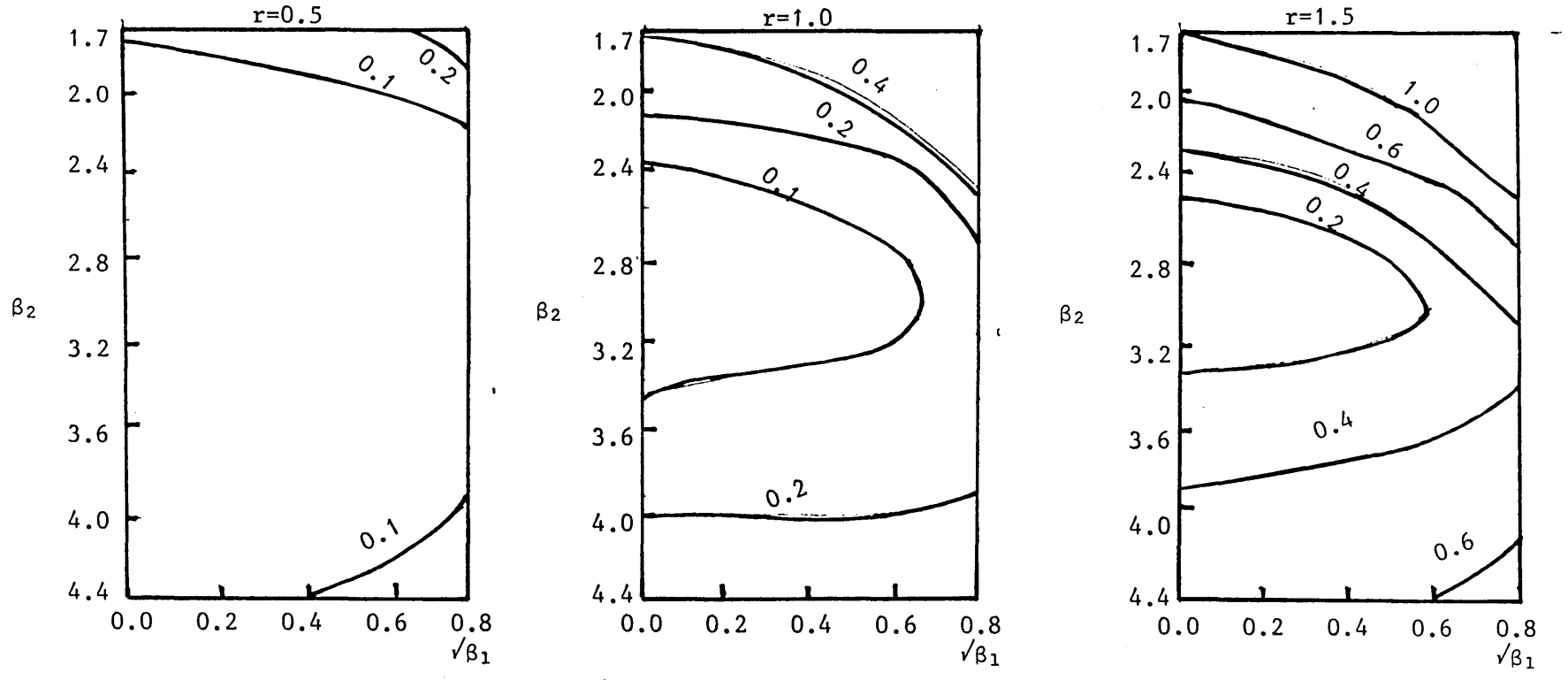


Figure 2.2.13 Contour diagrams for the maximum bias in β_2 (B_4) for $\sqrt{\beta_1}$ and β_2



2.3 Bivariate Distributions

This section shows how some of the results of section (2.2) may be extended to the bivariate case. The moments of the rounded distribution are obtained via the CF. Special consideration is given to the implications of rounding on the joint first moment of a bivariate normal distribution.

In communication engineering the quantization of signals from bivariate distributions was first investigated by Widrow (1961) and Watts (1961). Watts obtained the CF of quantized signals from a bivariate distribution by the same approach as he used for the univariate. Section (2.3.1) will show how the CF for the bivariate distribution can be derived from Watts' result. While he gave no proof for his CF result, a simple proof will be obtained for rounded distributions. Widrow (1961) considered the joint first moment of quantized signals for a bivariate normal distribution. He obtained an approximation to the bias caused by rounding in this moment. However his approximation is only suitable for lattice position $c = 0$ and is in fact incorrect. Using the CF for the rounded distribution, for the first time an exact expression is derived for the joint first moment for any bivariate distribution. It is shown how this expression can be further simplified if the CF exhibits certain symmetric properties.

The statistical literature has customarily assumed that Sheppard's corrections are suitable for finding the relationship between the moments of unrounded and rounded data from a bivariate distribution. For the bivariate normal it will be shown how Sheppard's corrections applied to the joint first moment may break down when there is high correlation. The implications of this on the correlation coefficient will be demonstrated.

2.3.1 Characteristic Function and Moments of a Rounded Bivariate Distribution

Let the two dimensional random variable (X,Y) have a bivariate distribution f(x,y). The CF of (X,Y) will be denoted by

$$\varphi_{XY}(t_1,t_2) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{it_1x+it_2y} f(x,y)dxy$$

Theorem 2.2

Let (X,Y) be a two dimensional random variable with CF $\varphi_{XY}(t_1,t_2)$ values of X and Y are rounded respectively into rounding lattice with intervals of width w_1, w_2 and lattice positions c_1, c_2. The result is the two dimensional random variable (X_R,Y_R). The CF of (X_R,Y_R) is given by

$$\varphi_{XY_R}(t_1,t_2) = \sum_{k=-\infty}^{+\infty}\sum_{\ell=-\infty}^{+\infty} e^{-i2\pi(kc_1+\ell c_2)}\varphi_{XY}\left[t_1 + \frac{2\pi k}{w_1}, t_2 + \frac{2\pi\ell}{w_2}\right] \times \frac{\mathrm{Sin}[\frac{1}{2}(t_1w_1+2\pi k)]\,\mathrm{Sin}[\frac{1}{2}(t_2w_2+2\pi\ell)]}{\frac{1}{2}(t_1w_1+2\pi k)\,\frac{1}{2}(t_2w_2+2\pi\ell)} \qquad (2.3\text{-}1)$$

Proof of Theorem 2.2

Using the convolution theorem and Poisson summation formula for two dimensions, the method of proving theorem (2.2) is similar to that given for the univariate distribution in section (2.1).

By letting the gain and shift in the bivariate quantizer system be (w_1,w_2) and (c_1,c_2) respectively, (2.3–1) can be obtained from Watts (1961). Watts provided no proof for the CF of the bivariate quantizer system. However, as indicated above, the proof for the CF of (X_R,Y_R) can be straight forward. Although a similar result to (2.4–1) has been used in quantization theory, this is the first time the parallel result has been applied to a rounded bivariate distribution.

The CF (2.3–1) can be used to determine the moments of (X_R,Y_R). However it is useful only for joint moments, as the moments for the marginal distributions of X_R and Y_R can be obtained from expressions for moments in the univariate case. Of particular importance in multivariate analysis is how much the joint first moment between two variables may be affected by rounding.

Corollary 2.2 to Theorem 2.2

Let (X,Y) be a two dimensional random variable with CF $\varphi_{XY}(t_1,t_2)$. Values of X and Y are rounded respectively into a rounding lattice with intervals of width w_1, w_2 and lattice positions c_1, c_2 the result is the two dimensional random variable (X_R,Y_R)

$$E[X_R Y_R] = E[XY] - \frac{w_1}{2\pi} \sum_{k=-\infty}^{\infty}{}' e^{-i2\pi k c_1} \frac{(-1)^k}{k} \varphi'_{XY}\left[\frac{2\pi k}{w_1}, 0\right]$$

$$- \frac{w_2}{2\pi} \sum_{\ell=-\infty}^{+\infty}{}' e^{-i2\pi \ell c_2} \frac{(-1)^\ell}{\ell} \varphi'_{XY}\left[0, \frac{2\pi \ell}{w_2}\right]$$

$$- \frac{w_1 w_2}{4\pi^2} \sum_{k=-\infty}^{+\infty} \sum_{\ell=-\infty}^{+\infty}{}' e^{-i2\pi(kc_1+\ell c_2)} \frac{(-1)^{k+\ell}}{k\ell} \varphi_{XY}\left[\frac{2\pi k}{w_1}, \frac{2\pi \ell}{w_2}\right] \qquad (2.3\text{-}2)$$

where $\varphi'_{XY}\left[\frac{2\pi k}{w_1}, 0\right] = \frac{d}{dt_2}\left[\varphi_{XY}(t_1, t_2)\right]_{\substack{t_1 = \frac{2\pi k}{w_1} \\ t_2 = 0}}$

$$\varphi'_{XY}\left[0, \frac{2\pi k}{w_2}\right] = \frac{d}{dt_1}\left[\varphi_{XY}(t_1, t_2)\right]_{\substack{t_1 = 0 \\ t_2 = \frac{2\pi k}{w_2}}}$$

$\sum'$ denoting summation excluding the zero term.

Proof

Due to the lengthy manipulation involved, only an outline proof will be given.

Partitioning the CF (2.3–1) into four parts we have

$$\varphi_{XY_R}(t_1, t_2) = \varphi_{XY}(t_1, t_2) \frac{\text{Sin } \frac{t_1 w_1}{2} \text{ Sin } \frac{t_2 w_2}{2}}{\left[\frac{t_1 w_2}{2}\right]\left[\frac{t_2 w_2}{2}\right]} \quad \begin{bmatrix}\text{When } k = \ell = 0. \\ \text{Let this term} \\ \text{be } A(t_1, t_2)\end{bmatrix}$$

$$+ \sum_{\ell=-\infty}^{\infty}{}' e^{-i2\pi \ell c_2} \varphi_{XY}\left[t_1, t_2 + \frac{2\pi \ell}{w_2}\right] \frac{\text{Sin } \frac{t_1 w_1}{2}}{\frac{t_1 w_1}{2}} \frac{\text{Sin } \frac{1}{2}(t_2 w_2 + 2\pi \ell)}{\frac{1}{2}(t_2 w_2 + 2\pi \ell)} \quad \begin{bmatrix}\text{when } k=0, \\ \ell \neq 0. \text{ Let} \\ \text{this term be} \\ B(t_1, t_2)\end{bmatrix}$$

$$+ \sum_{k=-\infty}^{+\infty}{}' e^{-i2\pi kc_1} \varphi_{XY}\left[t_1 + \frac{2\pi k}{w_1}, t_2\right] \frac{\text{Sin } \frac{1}{2}(t_1 w_1 + 2\pi k)}{\frac{1}{2}(t_1 w_1 + 2\pi k)} \frac{\text{Sin } \frac{t_2 w_2}{2}}{\frac{t_2 w_2}{2}} \left[\begin{array}{l}\text{when } k \neq 0, \\ \ell = 0. \text{ Let} \\ \text{this term be} \\ C(t_1, t_2)\end{array}\right]$$

$$+ \sum_{k=-\infty}^{+\infty}{}' \sum_{\ell=-\infty}^{+\infty}{}' e^{-i2\pi(kc_1 + \ell c_2)} \varphi_{XY}\left[t_1 + \frac{2\pi k}{w_1}, t_2 + \frac{2\pi \ell}{w_2}\right]$$

$$\frac{\text{Sin } \frac{1}{2}(t_1 w_1 + 2\pi k)}{\frac{1}{2}(t_1 w_1 + 2\pi k)} \frac{\text{Sin } \frac{1}{2}(t_2 w_2 + 2\pi \ell)}{\frac{1}{2}(t_2 w_2 + 2\pi \ell)} \left[\begin{array}{l}\text{When } k \text{ or } \ell \neq 0 \\ \text{Let this term be} \\ D(t_1, t_2)\end{array}\right]$$

We have, changing order of differentiation and summation

$$\left[\frac{d^2 A(t_1, t_2)}{dt_1 t_2}\right]_{t_1 = t_2 = 0} = - E[XY]$$

$$\left[\frac{d^2 B(t_1, t_2)}{dt_1 t_2}\right]_{t_1 = t_2 = 0} = \sum_{\ell=-\infty}^{+\infty}{}' e^{-i2\pi \ell c_2} \frac{(-1)^\ell}{\ell} \varphi'_{XY}\left[0, \frac{2\pi \ell}{w_2}\right] \qquad (2)$$

$$\left[\frac{d^2 C(t_1, t_2)}{dt_1 t_2}\right]_{t_1 = t_2 = 0} = \sum_{k=-\infty}^{+\infty}{}' e^{-i2\pi k c_1} \frac{(-1)^k}{k} \varphi'_{XY}\left[\frac{2\pi k}{w_1}, 0\right] \qquad (3)$$

$$\left[\frac{d^2 D(t_1, t_2)}{dt_1 t_2}\right]_{t_1 = t_2} = \sum_{k=-\infty}^{+\infty}{}' \sum_{\ell=-\infty}^{+\infty}{}' e^{-i2\pi(kc_1 + \ell c_2)} \frac{(-1)^{k+\ell}}{k\ell}$$

$$\times \varphi_{XY}\left[\frac{2\pi k}{w_1}, \frac{2\pi \ell}{w_2}\right] \qquad (4)$$

$$E[X_R Y_R] = (-i)^2 \left[\frac{d^2 \varphi_{XY_R}(t_1, t_2)}{dt_1 t_2}\right]_{t_1 = t_2} = - \left[(1)+(2)+(3)+(4)\right]$$

the required result.

If the CF has the following symmetric property

$$\varphi_{XY}(t_1, t_2) = \varphi_{XY}(-t_1, t_2) \qquad \text{[ie CF real]}$$

the result given in Theorem (2.2) can be further simplified to

$$E[X_R Y_R] = E[XY] - \frac{w_1}{2\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \varphi'_{XY}\left[\frac{2\pi k}{w_1}, 0\right] \cos(2\pi k c_1) \qquad (2.3\text{-}3)$$

$$- \frac{w_2}{2\pi} \sum_{\ell=1}^{\infty} \frac{(-1)^\ell}{\ell} \varphi'_{XY}\left[0, \frac{2\pi \ell}{w_2}\right] \cos(2\pi \ell c_2)$$

$$- \frac{w_1 w_2}{2\pi^2} \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \frac{(-1)^{k+\ell}}{k\ell} \left[\varphi_{XY}\left[\frac{2\pi k}{w_1}, \frac{2\pi \ell}{w_2}\right] \cos[2\pi(kc_1 + \ell c_2)]\right.$$

$$\left. + \varphi_{XY}\left[\frac{2\pi k}{w_1}, -\frac{2\pi \ell}{w_2}\right] \cos[2\pi(kc_1 - \ell c_2)]\right]$$

2.3.2 Bivariate Normal

Let the two dimensional random variable (X,Y) have a bivariate normal distribution with the following p.d.f.

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp(-\tfrac{1}{2}Q) \qquad \begin{matrix} -\infty < x < \infty \\ -\infty < y < \infty \end{matrix} \qquad (2.3\text{-}4)$$

where

$$Q = \frac{1}{(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]$$

$$-1 < \rho < 1, \quad \sigma_X > 0, \quad \sigma_Y > 0, \quad -\infty < \mu_X < -\infty, \quad -\infty < \mu_Y < \infty$$

In this section it will be assumed that $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = \sigma^2$. We lose no generality in the results by making such an assumption. Usually values of X and Y will be rounded corresponding to the same rounding lattice, ie $w_1 = w_2 = w$ and $c_1 = c_2 = c$. This is the most likely situation under rounding

and the only case considered in this section.

The CF of (2.3–4) where $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = \sigma$ is

$$\varphi_{XY}(t_1,t_2) = \exp\left[-\frac{\sigma^2}{2}(t_1^2+2\rho t_1 t_2+t_2^2\right]$$

As $\varphi_{XY}(t_1,t_2) = \varphi_{XY}(-t_1,-t_2)$ the joint first moment of (X_R,Y_R) can be obtained from (2.3–3). If values of (X,Y) are rounded into a rounding lattice with intervals of width w and lattice position c, then $E[X_RY_R]$ from (2.4–3) is

$$\begin{aligned} E[X_RY_R] = E[XY] &+ 4\rho\sigma^2 \sum_{k=1}^{\infty} (-1)^k \exp\left[-\frac{2\pi^2k^2}{r^2}\right]\mathrm{Cos}(2\pi kc) \\ &- \frac{r^2\sigma^2}{2\pi^2} \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty} \frac{(-1)^{k+\ell}}{k\ell}\left[\exp\left[-\frac{2\pi^2}{r^2}(k^2+\ell^2+2\rho k\ell)\right]\right. \\ &\times \mathrm{Cos}[2\pi c(k+\ell)] + \left.\exp\left[-\frac{2\pi^2}{r^2}(k^2+\ell^2-2\rho k\ell)\right]\mathrm{Cos}2\pi c(k-\ell)\right] \end{aligned} \qquad (2.3\text{–}5)$$

The main importance of (2.3–5) is that it allows us to determine how the joint first moment is distorted by rounding. This can be illustrated by examining the rounding bias in this moment relative to σ^2

$$B = \frac{E[X_RY_R] - E[XY]}{\sigma^2} \qquad (2.3\text{–}6)$$

This bias B can be obtained exactly from (2.3–5). In the literature two approximations to this bias have been used.

(a) In communication engineering, Widrow (1956) suggested that the bias B is given very closely by the approximation

$$B_W = \frac{r^2}{12} \exp\left[-(1-\rho)\frac{4\pi^2}{r^2}\right] \qquad (2.3\text{-}7)$$

However this approximation should be treated with caution for two reasons. Firstly at arriving at (2.3–7) Widrow assumed a lattice position c = 0. Thus the approximation does not reflect the possible effect of the lattice. Secondly the approximation was derived incorrectly. This was pointed out by Watts (1961). In arriving at his approximation Widrow missed out the term

$$4\rho\sigma^2 \sum_{k=1}^{\infty} (-1)^k \exp\left[-\frac{2\pi^2 k^2}{r^2}\right] \mathrm{Cos}(2\pi kc) \qquad (2.3\text{-}8)$$

The accuracy of Widrow's approximation will depend on whether (2.3–8) may be considered negligible and the extent to which the lattice effect may be ignored.

(b) In statistical literature the customary adjustment to the moments for rounding is that provided by Sheppard's correction. For the joint first moment we have

$$E[X_R Y_R] = E[XY] + \frac{r^2\sigma^2}{12} \qquad X = Y$$

$$= E[XY] \qquad X \neq Y$$

This implies that the bias B is $r^2/12$ when X = Y and zero elsewhere. As in the univariate case, Sheppard's correction will be strictly valid only if the CF vanishes outside a finite interval, ie for the bivariate case

$$\varphi_{XY}(t_1,t_2) = 0 \text{ for } |t| \geq \frac{2\pi}{w_1},\ |t_2| \geq \frac{2\pi}{w_2}$$

Widrow's approximation and Sheppard's correction have both been used in the past to estimate the bias caused by rounding in the joint first moment. No study has been made of the reliability of these two methods. The validity of these methods was investigated as follows.

A Fortran program was written to calculate B(2.3-6) and B_W(2.3-7). The rounding precision varied upto 2, for lattice positions c = -0.5,-0.45,...,0.5. Selected resutls for $\rho > 0$ are given in Table (2.3.1). Similar values of β were obtained for $\rho < 0$. Table (2.3.2) shows the bias given by Sheppard's method. As expected, the bias in the joint first moment increases as r and ρ increase in value. Widrow's approximation for $\rho < 1$ over estimates this bias, this being more marked for r > 1.0 and high correlation. The results show that Sheppard's method can be poor in estimating the rounding bias. This being especially so for high correlation and coarse rounding (r > 1.0).

The results indicate that Widrow's approximation gives a reasonable approximation to the bias in the joint first moment caused by rounding for r $\leq$ 1.0. For Sheppard's method this is only so when r $\leq$ 0.5.

Table 2.3.1

Bias B caused by rounding in joint first moment of a bivariate normal distribution. (Widrow's approximation B_W in brackets)

ρ	r = 2.0	r = 1.5	r = 1.0	r = 0.5
1.0	0.301-0.365 (0.333)	0.187-0.188 (0.188)	0.083 (0.083)	0.021 (0.021)
0.99	0.200-0.261 (0.302)	0.112-0.114 (0.157)	0.037 (0.056)	0.003 (0.004)
0.98	0.164-0.223 (0.274)	0.087-0.088 (0.132)	0.024 (0.038)	0.001 (0.001)
0.97	0.139-0.197 (0.248)	0.070-0.072 (0.111)	0.016 (0.025)	0 (0)
0.96	0.119-0.176 (0.225)	0.058-0.059 (0.093)	0.011 (0.018)	0 (0)
0.95	0.103-0.159 (0.203)	0.048-0.049 (0.078)	0.007 (0.012)	0 (0)
0.90	0.050-0.103 (0.124)	0.019-0.020 (0.032)	0.001 (0.002)	0 (0)
0.85	0.022-0.071 (0.076)	0.008-0.009 (0.013)	0 (0)	0 (0)
0.80	0.005-0.051 (0.046)	0.003-0.004 (0.006)	0 (0)	0 (0)
0.50	0 -0.016 (0.012)	0	0 (0)	0 (0)
0.10	0 -0.003 (0.001)	0	0 (0)	0 (0)

Note all values in table are multiples of σ^2

Table 2.3.2

Bias in joint first moment given by Sheppard's method

r	$\rho = 1$	$\rho \neq 1$
2	$0.3333\sigma^2$	0
1.5	$0.1875\sigma^2$	0
1.0	$0.0833\sigma^2$	0
0.5	$0.0208\sigma^2$	0

Although not common, one variable of a bivariate distribution may be subject to rounding. What implications may this have on the effect of rounding on the joint first moment? For a bivariate distribution in which only one variable is rounded, say X, the CF of the two dimensional rounded random variable (X_R,Y) is

$$\varphi_{XY_R}(t_1,t_2) = \sum_{k=-\infty}^{+\infty} e^{-i2\pi kc}\varphi_{XY}\left[t_1 + \frac{2\pi k}{w}, t_2\right] \frac{\text{Sin } \frac{1}{2}(tw+2\pi k)}{\frac{1}{2}(tw+2\pi k)} \quad (2.3\text{-}8)$$

From (2.3–8) it follows that the joint first moment of (X_R,Y) is given by:

$$E[X_RY] = E[XY] - \frac{w}{2\pi} \sum_{k=-\infty}^{+\infty}{}' e^{-i2\pi kc} \frac{(-1)^k}{k} \varphi'_{XY}\left[\frac{2\pi k}{w}, 0\right] \quad (2.3\text{-}9)$$

For the bivariate normal distribution (2.3–9) becomes

$$E[X_RY] = E[XY] - 2\sigma^2\rho \sum_{k=1}^{\infty} (-1)^k \exp\left[-\frac{2\pi^2k^2}{r^2}\right] \cos(2\pi kc) \quad (2.3\text{-}10)$$

where $r = w/\sigma$.

There is only one error term in (2.3–10), which can be very small even for coarse rounding. When only one variable has been rounded, the rounding bias in the joint first moment can be considerably reduced, as compared when both variables are rounded, this being more so for coarse rounding. This is illustrated by the results shown in Table (2.3.3) for $r = 1.5, 2.0$ at $c = 0.5$.

Table 2.3.3

Bias in joint first moment relative to σ^2, for bivariate normal distribution for r = 2.0, 1.5 at c = 0.5

ρ	One variable rounded		Both variables rounded	
	r = 1.5	r = 2.0	r = 1.5	r = 2.0
0.95	$2.94(10)^{-4}$	$1.37(10)^{-2}$	0.049	0.159
0.90	$2.79(10)^{-4}$	$1.29(10)^{-2}$	0.020	0.103
0.70	$2.17(10)^{-4}$	$1.01(10)^{-2}$	0.001	0.031
0.50	$1.55(10)^{-4}$	$7.19(10)^{-3}$	$3.1(10)^{-4}$	0.016
0.10	$3.10(10)^{-5}$	$1.44(10)^{-3}$	$1.2(10)^{-4}$	0.003

Population Correlation

The correlation for the rounded distribution of (X_R,Y_R) is given by

$$\rho_R = \frac{E[X_RY_R] - E[X_R]E[Y_R]}{\sqrt{V[X_R]V[Y_R]}} \qquad (2.3\text{-}11)$$

As we have already obtained expressions for the expectations and variances in (2.3–11) the value of ρ_R can be obtained. Generally rounding caused the correlation to decrease, ie $|\rho_R| < |\rho|$. The one exception is when ρ is unity. When X and Y are rounded according to the same lattice and $\rho = 1$, then $X_R = Y_R$. Thus from (2.3–11) ρ_R is also unity. This makes sense. When ρ is unity, (X,Y) lie on a line, rounding only rearranges them along the line. However, more generally, when X and Y are rounded according to a different rounding lattice, rounding will cause (X,Y) values to be displaced either side of the

line and make $\rho_R < 1$.

Figure (2.3.1) shows a plot of ρ against ρ_R for three values of r at c = 0.5. This is typical of the relationship found between ρ and ρ_R. As illustrated by Figure (2.3.1) the correlation is more affected by the rounding process for values of ρ between 0.95–0.5, this being more noticeable as the rounding becomes more coarse.

The reduction in correlation as a result of rounding has an interesting implication on the usual estimate of ρ. For unrounded data the estimate of ρ is

$$\hat{\rho} = \frac{\frac{1}{n}\sum_i (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{S_X^2 \times S_Y^2}}$$

where $S_X^2 = \sum_i (X_i-\bar{X})^2/n$ and $S_Y^2 = \sum_i (Y_i-\bar{Y})^2/n$

If no adjustment is applied to $\hat{\rho}$ for rounding, then $\hat{\rho}$ will tend to under estimate ρ. If Sheppard's corrections are applied to the variance, while leaving the covariance unaltered (ie zero correction) then $\hat{\rho}$ may exceed unity in situations of high correlation.

In section (2.3) the results for the univariate situation were extended to bivariate distributions. The expression for the CF of a rounded univariate distribution can be generalised to an n–dimensional rounded random variable. From this CF the moments of a rounded multivariate distribution can be obtained. However, this extension may not always be necessary. For example, consider the multivariate normal distribution. The most important joint moment is the joint moment about

two variables. The effects of rounding on this moment can be obtained from the bivariate result given by (2.3–5).

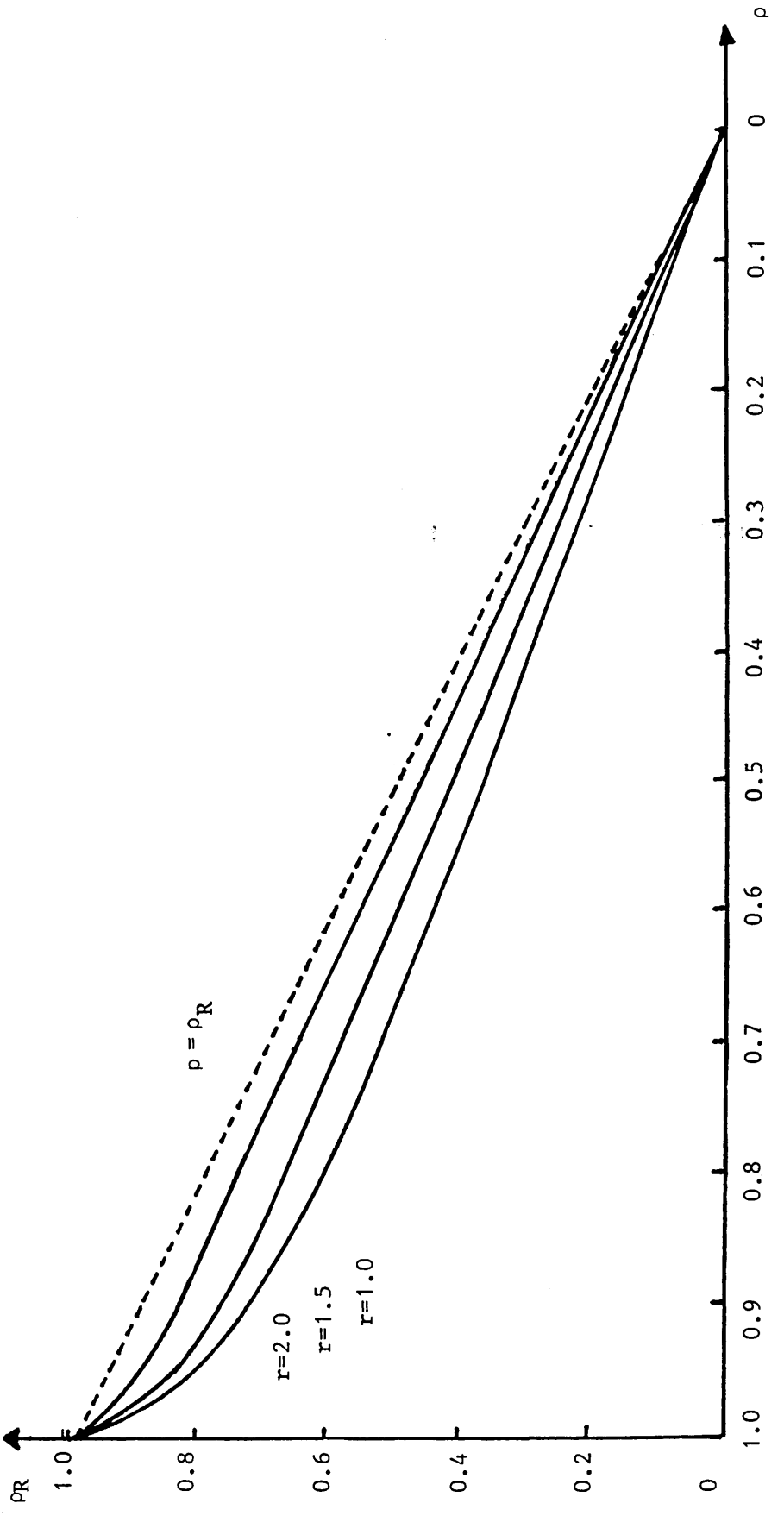


Figure 2.3.1 ρ vs ρ_R for r = 2.0, 1.5, 1.0 at c = 0.5

2.4 Conclusions

This chapter has given a general method for assessing the extent of the bias in the moments of a continuous distribution caused by rounding. This has been a more detailed study than any given in the past, as it has considered not only the degree of rounding, but also lattice position and shape of the distribution. Also the reliability of Sheppard's corrections have been considered.

For the normal distribution Sheppard's corrections were found to be a reasonable approximation to the moments of rounded data for $r \leq 2.0$. As illustrated by the gamma distribution, departure from normality may result in rounding causing a greater bias in the moments. This makes Sheppard's corrections less reliable. The mount of bias in the moments caused by rounding was found to be closely related to the shape of the gamma distribution. As it became increasingly less symmetrical, the bias increased. Previous work on precision of data has concentrated on the effect of the degree of precision of the rounded data (r) on the distribution. The results from the normal and gamma distributions suggest that this is not the only important factor. The position of the rounding lattice, and especially the shape of a distribution must be taken into account.

In section (2.2.3) the Johnson System of distributions was used to illustrate the relationship between the shape of a distribution and the bias in the four parameters μ, σ, $\sqrt{\beta_1}$ and β_2 caused by rounding. The results indicated the extent to which $\sqrt{\beta_1}$ and β_2 determined the rounding bias in these four parameters. The departure from symmetry was a crucial factor in deciding the size of the rounding bias. Generally as the distribution became increasingly non-symmetrical, the rounding bias increased. The importance of this rounding bias in the four parameters to the

statistician will depend upon the particular application. Usually values of r encountered in practice are often sufficiently small to render this bias negligible. However, as the results indicate, the value of r for which this may be so can vary depending on the shape of the distribution.

In section (2.3), the effect of rounding on the moments of a bivariate distribution was considered. Attention was focussed on the joint first moment. For the bivariate normal distribution, rounding bias in the joint first moment depends on the correlation. For fixed r, as the correlation increased so did the rounding bias in this moment. Sheppard's corrections were less reliable than Widrow's in approximating this bias. Both methods were a poor approximation for coarse rounding ($r > 1.0$).

Generally in the bivariate normal, rounding decreased the correlation between the variables. Where high correlation existed Sheppard's correction could be unreliable in adjusting the joint moment for rounding. This could lead to the correlation coefficient $\hat{\rho}$ exceeding unity. This demonstrated how Sheppard's corrections should be handled with care in the bivariate (multivariate) situation. Although only the bivariate normal was considered, the theory developed can be used to investigate the effect of rounding on other bivariate distributions.

CHAPTER 3

THE EFFECT OF ROUNDING ON THE SIGNIFICANCE LEVEL OF CERTAIN NORMAL TEST STATISTICS

3.1 Introduction

The underlying theory upon which many statistical tests are based, assumes that the variable or variables sampled are continuous. There is no such thing, in practice, as a continuous variable. It is often expedient for us to consider observations as being rounded from an underlying continuous distribution. To date, there has been very little research into the effect of rounding on a statistical test. This chapter investigates the performance of test statistics under rounding. We will be particularly interested in the degree of precision (r) to which a set of data should be recorded before applying a statistical test. There is considerable vagueness concerning what level of precision should be used. Most statisticians know, for example, that tests of means tend to be robust under departures from normality and that chi-squared and F tests of variance do not. However, they know little about what happens to the significance level and power when the data have been rounded. The reason for this lack of quantitative knowledge has been the absence of a careful accurate study of the effect of rounding on statistical tests. The absence of such a study is primarily due to the following:

(a) the problem of determining the exact distribution of the test statistic for rounded data;
(b) mathematical approximations that have been studied lack accuracy;
(c) Monte Carlo studies in the past required an exorbitant amount of computer time to achieve respectable precision in the results.

Because large computers are now available, studies by Monte Carlo methods offer an excellent approach to investigating the effect of rounding on test statistics. However, to date no one has published a study of this type.

Several authors have considered the problem of rounding and test statistics. Student (1908) gives a most interesting discussion into the possible problems of coarse rounding on statistical procedures. Student's experimental results suggest that the distribution of the single t statistic for rounded and unrounded data will be approximately the same if the sample size is large. Although Student made no detailed study of the performance of the t statistic under rounding, he was the first to point out the possible implications that rounding may have. Fisher (1936) advocated that Sheppard's corrections should be used for the purpose of estimation, but not usually for tests of significance. Eisenhart (1947) pointed out that use of a Sheppard's correction can make the t value imaginary, as the corrected estimate of the variance can be negative. Geddeback (1968) advocated that Sheppard's corrections should be avoided in the analysis of variance. Krutchoff (1967) states "There is no such thing, in practice, as a continuous random variable. It is often expedient for us to consider observations as being rounded from an underlying continuous random variable." He illustrates this point by showing how rounding can cause the F statistic to have a non-zero probability of a zero in the denominator and as such the mean of this statistic will not exist.

Eisenhart (1947) was the first to study in any detail how rounding affects statistical tests. He gave a set of rules, to the problem of how large a sample size n needs to be for a given w for judging the suitability of a particular coarseness of rounding when applying the one sample t-test, chi-squared test for a variance and F-test for equality of two variances. [Details in literature review]. His study has the following limitations. Eisenhart's recommendations were based on the probability of a sample variance obtained from the rounded data being zero. This gives no indication of the performance of the test statistics with respect to level of significance or power under rounding. His recommendations were based only on

samples as large as n equal to 7. In Preece (1982), text book examples of the paired t–test are examined with respect to the degree of precision of data recording. From these examples, he concludes that, for coarse rounding, the value obtained for a paired t–statistic depends crucially both on the rounding interval applied and the position of the rounding grid relative to the origin. As Preece points out, final conclusions cannot be drawn from several examples and further work on the effect of rounding on test statistics is called for. Riley, Bekele and Shrewsbury (1983) adopt a similar approach to Preece in investigating the possible effect of rounding on test statistics. To examine the effect that different degrees of precision have on the analysis of variance, they present several examples where data has been recorded initially to a good degree of precision. For each example the analysis of variance is obtained for various degrees of rounding. From this small set of examples they make some general points about the effect of rounding on the mean squares. The main finding can be summarised as follows. As rounding became more and more severe the mean squares began to behave very erractically. However data could be rounded appreciably before loss of information became significant. With respect to the various recommendations to what degree of precision should be used on rounded data, they concluded that Dyke's rule (Dyke, 1974 pp163–164) gave a safe degree of precision for every set of data they examined.

The investigations by Preece (1982) and Riley, Berkele and Shrewsbury (1983) have a major limitation. They consisted of looking at the effect of rounding on specific examples. The actual distribution of the test statistic for rounded data was not obtained. As a result no general conclusions could be established about the performance of a test statistic for rounded data. A study which involved the probability distribution of the test statistic under rounding would enable significance

level and power to be considered. Such a study would be of value in supplying answers about how robust specific test statistics are for rounded data. A problem in producing such a study, however, is the very large amount of computer time required. Either one must find a way to find this time or a way of reducing the amount of time required without decreasing the quality of the study. The study undertaken in this chapter does both through the development of purpose written programs which reduce the required computer time to "only a large amount" and through the Polytechnic computer service support via low priority computer use over a long period of time.

The objective of the present extensive study is to precisely quantify the significance level and power of statistical tests on rounded data over many distributions. The study has two sections, namely when the parent population is normal or non-normal. Chapters 3 and 4 respectively considers the significance level and power levels of these tests when data comes from a rounded normal distribution. Chapter 5 deals with both significance level and power for a selection of non-normal rounded distributions.

3.2 Description of the Investigation

This chapter is concerned with the effect of rounded normal data on the significance level of a test. Many statistical tests could have been investigated. It was decided to investigate test statistics which are frequently used in practice, these being the one sample t-test, the chi-squared test for variance, the two sample t-test, F-test for equality of variances and F-test in the one and two way analysis of variance. Choosing such a selection of tests allowed a wide coverage of the possible implications of rounding on test procedures.

The main distortion caused by rounding is the discreteness it introduces into the sampling distribution of the test statistic. Although the moments may not be widely affected, the area in the tails of the sampling distribution may be changed. Examining the moments of the test statistic under rounding will indicate the possible effect of rounding. However evaluation of the exact distribution of the sampling distribution of the test statistic under rounding is required for a detailed examination of the possible changes in the tails of the sampling distribution.

The following approaches were used to examine the implications of rounding on the significance level of a test:

(i) Approximations to the sampling moments of the test statistics. These theoretical results have some bearing on the distribution of the test statistics in sampling from rounded normal populations. However they will provide only a rough outline of what characteristics are to be expected when sampling from rounded normal populations, they do not supply answers in numerical terms of the effect of rounding on the significance level of a test. This is why the exact distribution of the sampling distribution of the test statistic is required.

(ii) The exact distribution of the test statistic for rounded data was obtained. By constructing all sample configurations, the exact distribution of the test statistic for rounded data can be obtained. This method was used for small sample sizes. However, it became uneconomical to use this method for large samples and for the analysis of variance.

(iii) The sampling distribution of the test statistic for rounded data was obtained by Monte Carlo methods. Simulation was used where it was impractical in terms of computer time to obtain the exact distribution in

(ii).

Two Fortran programs were written for the necessary analysis. The program EXACT generated every possible sample of size n from a normal population that had been rounded according to a lattice with rounding interval w and lattice position c. The required test statistic was calculated and the percentage of samples where the designated statistic fell about or below the α significance level limits for normal theory conditions was recorded.

The program SIMUL generates N random samples of size n from a normal population which has been rounded to the specific w and c. As in the program EXACT, from each rounded sample the required statistics are calculated and the percentage of samples where the designated statistic fell about or below the α significance level limits for normal theory conditions was recorded. Both the EXACT and SIMUL programs gave the mean and variance of the test statistic for rounded data.

For this study the significance level of the test statistic under rounding was evaluated for values corresponding to the lower and upper 0.1%, 1.0%, 2.5% and 5% points under normal theory conditions, with no rounding. This range of significance levels allowed us to cover one tailed tests at α = 0.001, 0.01, 0.05 and two tailed tests at α = 0.05. Sample sizes from 2 to 25 were considered for the one and two sample t-tests, chi-squared test and F-test. As the sample size increased in size the discreteness in the sampling distribution of the test statistic caused by rounding had less effect. As a result a sample size of 25 was found to give in most situations a good indication of the effect of rounding for larger sample sizes. Where this was not so, sample sizes larger than 25 were considered.

For the one and two way analysis of variance various levels of the factor(s) were considered. The degree of precision ranged upto 2 and lattice positions c = −0.5, −0.4, ..., 0.4, 0.5 were used. A value of r beyond 2 is extremely coarse rounding and is impractical in most situations.

The results from the simulation were based on 100,000 iterations. That is, 100,000 values of each test statistic were generated for estimating each significance level under rounding. This number of iterations was necessary for respectable precision, especially for the 0.1% level of significance. Of course the results obtained for the significance levels by simulation are subject to sampling errors. For simulations of 100,000 iterations these will be small. For example, the standard error of our estimates of the significance level will be $6.89(10)^{-4}$ for $\alpha = 0.05$ and $3.15(10)^{-4}$ for $\alpha = 0.01$, by simple binomial calculations.

Quality of Results

Both the EXACT and SIMUL programs were tested to check the validity of their results. For example, an independent check on the results given by SIMUL program was provided by obtaining the significance levels for the test statistics when the normal population was subject to no rounding. They were found to be in very close agreement with the expected results. An independent check of the EXACT program was provided by comparing the results with those obtained manually. Of course this was only possible for small sample sizes (n=2,3). A final check was established by comparing the results for significance levels obtained from both EXACT and SIMUL programs. They were found to be in very close agreement.

3.3 Test Statistics

In this section it is assumed that the normal distributions have mean zero and variance one. This is convenient as in this situation r = w and the effect of the position of the distribution on the rounding lattice is simply given by the value of c. By using a standardised normal we lose no generality in the results. Throughout this section α will denote the level of significance of the test for samples drawn from a normal population subject to no rounding, while α_R will be the resulting level of significance of the test where the samples have been drawn from a rounded normal population. α_R is simply the probability that for rounded data, the test statistic fell above or below the α significance level limits. α_R may be regarded as the 'true' level of significance for normal rounded data.

The results are presented as follows. For each test statistic:

(i) Before discussing the results, approximate expressions are given for the moments of the test statistics for a normal population that has been subject to rounding. These approximate moments help to indicate any possible changes in the distribution of the test statistic under rounding.

(ii) Results from EXACT and SIMUL programs will be discussed. The discussion for convenience is divided into three sections according to sample size n, these being for $n \leq 5$, $n = 10$, and $n = 25$.

(iii) Finally a table is given which provides the values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25. When using a single or two tailed test, r is acceptable if the $\alpha\%$ significance level for unrounded normal data is:

(a) 5%, while for rounded data with degree of precision r, α_R lies between 4%–6%.

(b) 1%, while for rounded data with degree of precision r, α_R lies between 0.5%–1.5%.

(c) 0.1%, while for rounded data with degree of precision r, α_R lies between 0.0%–0.2%.

The results indicate the extent to which the test statistics are affected by rounding. The object of providing recommended values of (n,r) in which the level of significance is within certain bounds is not to limit the loss of information caused by rounding, but to indicate the circumstances under which rounded data can be validly analysed while keeping the level of significance within reasonable bounds. Of course in some situations the decision whether to reject H_0 or not will be different for rounded or unrounded data. Rounding may cause the test statistic to change from being significant to not significant or vice versa. However, overall, the significance level of the test will remain within acceptable bounds for the recommended values of (n,r).

At the end of the chapter the degree of precision r that may be regarded acceptable for values of n besides 5, 10, and 25 is given. Appendix B gives a list of all the output produced by the EXACT and SIMUL programs. This appendix also contains tables of results that are referred to in this chapter.

3.3.1 One sample t–test

Let $\underline{X} = (X_1,\ldots,X_n)$ be a random sample of size n from a normal population X. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the rounded value

X_i corresponding to a rounding lattice with interval of width w and lattice position c. For testing the hypothesis $H_0:\mu = \mu_0$ the t-test statistic is given by:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \qquad X \sim N[\mu_0, \sigma^2] \qquad (3.3\text{-}1)$$

and under rounding

$$t_R = \frac{\bar{X}_R - \mu_0}{S_R/\sqrt{n}} \qquad (3.3\text{-}2)$$

where

$$\bar{X} = \sum_i \frac{X_i}{n}, \quad S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2, \quad \bar{X}_R = \sum_i \frac{X_{Ri}}{n}, \quad S_R^2 = \frac{1}{n-1} \sum_i (X_{Ri} - \bar{X}_R)^2$$

As we have assumed that $X \sim N[0,1]$, (3.3-2) becomes

$$t_R = \frac{\bar{X}_R}{S_R/\sqrt{n}} \qquad \text{where } r = w$$

To obtain the approximation to the moments of t_R, we can use the work of Geary (1947), where he found expansions for the moments of a one sample t-test statistic, for samples drawn from non-normal populations. From Geary's results:

$$\begin{aligned} E[t_R] &= -\frac{1}{2\sqrt{n}} \sqrt{\beta_{1R}} - O(n^{-3/2}) \\ V[t_R] &= 1 + \tfrac{1}{4}(8+7\beta_{1R})/n + O(n^{-2}) \\ \sqrt{\beta_1}(t_R) &= -2\sqrt{\beta_{1R}}/\sqrt{n} - O(n^{-3/2}) \\ \beta_2(t_R) &= 3 + 2(6-\beta_{2R}+6\beta_{1R})/n + O(n^{-2}) \end{aligned} \qquad (3.3\text{-}3)$$

In Chapter 2 expressions were obtained for finding the four parameters μ_R, σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R}. For example, by substituting the values of $\sqrt{\beta_{1R}}$ and β_{2R} into (3.3-3) the moments of t_R can be obtained for a given r and c, when the sample size is large. However we only require the moments of a test statistic where n is large, to provide a rough outline of what to expect when sampling from a rounded normal population. Thus an approximation to the parameters μ_R, σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R} will be suitable. In Chapter 2 Sheppard's corrections were found to give a reasonable approximation to these parameters for $r \leq 2$ when the distribution is normal. From (2.2-5) and assuming a standard normal distribution we have:

$$\begin{aligned} \mu_R &= 0 \\ \sigma^2_R &= 1 + r^2/12 \\ \sqrt{\beta_{1R}} &= 0 \\ \beta_{2R} &= \left[3 + \frac{r^2}{2} + \frac{r^4}{80}\right] / \left[1 + \frac{r^2}{12}\right]^2 \end{aligned} \qquad (3.3\text{-}4)$$

Hence using (3.3-3) and (3.3-4) approximations to the moments of t_R, when X has a standard normal distribution, are given by:

$$\begin{aligned} E[t_R] &= 0 - O(n^{-3/2}) \\ V[t_R] &= 1 + \frac{2}{n} + O(n^{-2}) \\ \sqrt{\beta_1}(t_R) &= 0 - O(n^{-3/2}) \\ \sqrt{\beta_2}(t_R) &= 3 + \frac{2}{n}\left[3 + \frac{r^4}{120} / \left[1 + \frac{r^2}{12}\right]^2\right] + O(n^{-2}) \end{aligned} \qquad (3.3\text{-}5)$$

If the normal population is subject to no rounding we have:

$$E[t] = 0$$

$$V[t] = \frac{n-1}{n-3} = 1 + \frac{2}{n} + O(n^{-2}) \qquad (3.3\text{-}6)$$

$$\sqrt{\beta_1}(t) = 0$$

$$\sqrt{\beta_2}(t) = \frac{3(n-3)}{(n-5)} = 3 + \frac{6}{n} + O(n^{-2})$$

Study of the moments of t_R (3.3–5) and t (3.3–6) shows that in terms of order n^{-1}, rounding affects only the kurtosis, and that the effect is clearly negligible for $r \leq 2.0$. These moment results suggest that the distribution of t and t_R are similar. However the moment results are for large n and the discontinuities in t_R have a serious effect on the significance level of the test for small n.

EXACT/SIMUL Results

$n \leq 5$

In Appendix B, Table (B.1) shows the range in values of α_R for n = 5, where r = 2.0, 1.0 and 0.5.

Like the distributions of all functions of rounded observations, the distribution of t_R is discontinuous. Generally for a given r, as n decreases in size the discontinuities in t_R become more numerous in any given interval and the steps increase in size. A further complication as n becomes small is that the probability of $S^2_R = 0$ increases. This causes t_R to be either $+\infty$ or $-\infty$, according to the sign of the numerator of (3.3–2). The exception is where $\bar{X}_R$ equals the population mean, when $t_R = 0/0$, which can be defined to be equal to zero. This seems a sensible definition as 0/0 indicates that the X_{Ri} are all the same and are equal to the

population mean and under H_0 we would expect $t_R = 0$. Table (3.3.1) shows that values $t_R = +\infty$ or 0/0 occur with annoying frequency in very small samples, especially for coarse rounding ($r > 1$). The value of c is important in determining the probability of such values of t_R for $r > 1.0$. For such small values of n, the infinite values of t_R caused an inbalance in the values of α_R between the upper and lower tails. As expected increasing n or decreasing r caused the inbalance to become low. As shown by the t_R values for $n = 5$ in Table (B.1), this inbalance is very noticeable for $r = 2.0$.

Table 3.3.1

Probability that $t_R = \pm\infty$ or 0/0 for samples of size n drawn from a rounded normal population in a single sample t–test.

r	2		1.5		1.0	0.5	0.25
n \ c	0	0.5	0	0.5	all c	all c	all c
2	0.511	0.457	0.391	0.384	0.271	0.140	0.070
3	0.326	0.271	0.183	0.163	0.085	0.022	0.006
4	0.218	0.104	0.094	0.071	0.028	0.004	0.000
5	0.148	0.049	0.051	0.031	0.009	0.001	
6	0.125	0.024	0.028	0.013	0.003	0.000	
7	0.069	0.011	0.015	0.006	0.001		
8	0.047	0.005	0.008	0.002	0.000		

Note: The probabilities in Table (5.3.1) are also the probability that S^2_R is equal to zero.

For small n, t_R will have a large number of discontinuities, especially for coarse rounding. This will result in the distribution of t_R being a poor approximation to the distribution of t. We would expect the α and α_R values to be considerably different unless the value of r is low. The results for α_R confirm this. For example, as shown by the results in Table (B.1), α and α_R are only in close agreement for n = 5, where $r \leq 0.5$.

A clear pattern emerged in the values of α_R, which were caused by the influence of the lattice. For c in the range [−0.5,0), the lower tail values of α_R are the same as the upper tail values, for c in the range (0,+0.5]. For c = 0 and ± 0.5 the upper and lower tail values of α_R are the same. An example of this pattern is seen in Table (B.1). For the one sample t−test, this pattern will always occur when the unrounded distribution is symmetrical.

n = 10

Table (B.2) shows the range in values of α_R for n = 10, where r = 2.0, 1.5, 1.0 and 0.5.

For this size sample values of $t_R = \pm\infty$ or 0/0 are no real problem. Only for $r \geq 1.5$ is there a strong disagreement between the α and α_R values.

n = 25

Table (B.3) shows the range in values of α_R for n = 25, where r = 2.0 and 1.5. Also given in the corresponding range in the mean and variance of t_R. Even for rounding as coarse as r = 2.0, the mean and variance of t and t_R were found to

be very similar. This confirmed the results from the moments (3.3–5), which indicated the mean and variance of t_R will be very close to those for t for large n and $r \leq 2.0$. For n as large as 25, the distribution of t_R will closely approximate that of t, even for coarse rounding. This is evident from the results in Table (B.3) where for practical purposes there is very little difference between the values of α_R and α for rounding as coarse as r = 1.5.

Table (3.3.2) gives the values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25 (definition of 'acceptable' in section 3.3). In this table, the column 0.1/1.0/5.0 gives the range of r which is acceptable for these three levels of significance. Column 0.1/5.0 gives the range of r which is acceptable at both these levels of significance. Column 5.0 gives the range of r which is acceptable for just this level of significance.

Table 3.3.2

The values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25 in a one sample t–test.

	One tailed test						Two tailed test
α(%)	0.1/1.0/5.0		1.0/5.0		5.0		5.0
n	LT	UT	LT	UT	LT	UT	
5	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$
10	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$
25	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$

Note: LT = lower tail UT = upper tail

Immediately noticeable from Table (3.3.2) is that there is no difference in the ranges of r between lower and upper tails. As n increases in size, the degree of precision of the data can be decreased without any further deterioration in the significance level.

3.3.2 Chi-squared test for variance

Let $\underline{X} = (X_1,\ldots,X_n)$ be a random sample of size n from a normal population X. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the rounded value of X_i corresponding to a rounding lattice with interval of width w and lattice position c. For testing the hypothesis $H_0:\sigma^2 = \sigma_0^2$, the chi-squared test statistic is given by:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \qquad (3.3\text{-}7)$$

and under rounding

$$\chi_R^2 = \frac{(n-1)S_R^2}{\sigma_0^2} \qquad (3.3\text{-}8)$$

where S^2 and S^2_R are defined as in (3.3-2).

As we have assumed that $X \sim N[0,1]$, (3.3-8) becomes

$$\chi_R^2 = (n-1)S_R^2 \quad \text{and} \quad r = w \qquad (3.3\text{-}9)$$

To obtain the moments of X^2_R we can use the exact first and second moments of S^2 in terms of the moments of the population from which the sample has been drawn (Church, 1925).

$$E[S_R^2] = \sigma_R^2 \qquad (3.3\text{-}10)$$

$$V[S_R^2] = \left[\frac{\beta_{2R}-3}{n} + \frac{2}{n-1}\right]\sigma_R^4$$

Using (3.3–10) and Sheppard's corrections (3.3–4), approximations to the first two moments of χ^2_R, where X has a standard normal, are given by

$$E[\chi_R^2] \simeq \left[1 + \frac{r^2}{12}\right](n-1) \qquad (3.3\text{-}11)$$

$$V[\chi_R^2] \simeq \left[1 + \frac{r^2}{12}\right]^2\left[2(n-1) - \left[1 - \frac{1}{n}\right]^2\left[\frac{r^4}{120} \Big/ \left[1 + \frac{r^2}{12}\right]^2\right]\right]$$

If the normal population is not subject to rounding we have

$$E[\chi^2] = n-1 \quad , \quad V[\chi^2] = 2(n-1) \qquad (3.3\text{-}12)$$

Study of the moments of χ^2_R (3.3–11) and χ^2 (3.3–12) shows that the controlling factor is $(1+r^2/12)$, which will never vanish no matter what the size of n. Of course the shape of the distribution of χ^2 changes under rounding, but it will be the increased mean and variance which will have the greatest effect on the significance levels. The increased mean will cause the distribution of χ^2 to shift to the right, resulting in the α_R values in the lower tail being generally less than those in the upper tail.

EXACT/SIMUL Results

$n \leq 5$

Table (B.4) shows the range in values of α_R for n = 5, where r = 2.0, 1.0, 0.5 and 0.25.

For values of $n \leq 5$, a major problem caused by rounding is that there is a high probability that $\chi^2_R = 0$, as the probability of $S^2_R = 0$ is high. [Table(3.3.1)]. For very small n (n = 2 or 3), or coarse rounding ($r > 1.0$), this caused a high concentration of probability at zero and resulted in the lower tail values of α_R being greater than the upper tail values. The values of α_R for n = 5 where r = 2.0 given in Table (B.4) are typical of the type pattern found for coarse rounding.

The first moment of χ^2_R indicated that rounding would cause a shift to the right in the distribution of the test statistic. Generally, the influence of this shift in the distribution was not clear until the probability of χ^2_R being zero was low. For example at n = 5, the effect of this shift in the distribution on the values of α_R was not noticeable until $r \leq 1.0$ [Table (B.4)]. The probability of χ^2_R being zero can cause the lower tail significant levels to have identical values for $r \geq 1.0$.

The results indicated that the values of α_R and α would only be in close agreement if the data had been recorded to a high degree of precision. As shown by the results in Table (B.4) for n = 5, this was so for $r \leq 0.25$ only.

n = 10 and 25

Tables (B.5) and (B.6) show the range in values of α_R respectively at n = 10 and 25, where r = 1.5, 1.0, 0.5 and 0.25.

For these sample sizes, values of $\chi^2_R = 0$ have no noticeable influence on the significance level of the test. The results from the moments (3.3-11) indicated that rounding will cause the mean and variance of χ^2 test statistic to increase by a factor $(1+r^2/12)$ and $(1+r^2/12)^2$ respectively. The increase in the mean has the greatest effect on the significance level of the test. As indicated by the results in Tables (B.5) and (B.6), this will cause the values of α_R in the lower tails to be less than those in the upper tail. For fixed n, as r increased the inbalance in the values of α_R between the two tails increased. However, more unusual is that for fixed r, as n increased the results showed a tendency for the inbalance between the two tails to increase also. This is because the amount by which rounding causes the distribution of the test statistic to shift to the right is dependent on the value of n. Table (3.3.3) illustrates this point for n = 10, 25 and 90, where r = 1.0.

Table (3.3.3)

Range of values of α_R for n = 10, 25 and 90 when r = 1.0

	α(%) lower tail				α(%) upper tail			
n	0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
10	0.06-0.09	0.44-0.60	2.05-2.47	3.32-3.33	6.72-6.75	3.74-3.77	1.75-1.77	0.23-0.23
25	0.05-0.07	0.51-0.63	1.43-1.72	2.88-3.17	8.74-9.39	5.04-5.34	2.23-2.35	0.30-0.32
90	0.01-0.02	0.24-0.27	0.67-0.76	1.57-1.64	13.95-14.23	8.15-8.31	3.99-4.12	0.60-0.68

As shown by the results in Tables (B.5) and (B.6), there was a strong agreement between values of α_R and α at n = 10 and 25 when $r \leq 0.25$. Table (3.3.4) gives the values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25.

Table 3.3.4

The values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25 in a chi-squared test for a variance.

	One tailed test						Two tailed test
α(%)	0.1/1.0/5.0		1.0/5.0		5.0		5.0
n	LT	UT	LT	UT	LT	UT	
5	$r \leq 0.5$	$r \leq 0.25$	$r \leq 0.5$	$r \leq 0.25$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$
10	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 1.0$
25	$r \leq 0.5$	$r \leq 0.25$	$r \leq 0.5$	$r \leq 0.25$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$

The results in Table (3.3.4) indicate how the chi-squared test for a variance is not very robust to rounding particularly in the upper tail. For the two tailed situation the effect of rounding seems to be less, as two tailed tests have a compensating factor between the upper and lower tails.

3.3.3 Two sample t-test

Let $\underline{X} = (X_1,\ldots,X_{n_1})$ and $\underline{Y} = (Y_1,\ldots,Y_{n_2})$ be independent random samples of sizes n_1 and n_2 from normal populations X and Y with means μ_X, μ_Y and

variances σ^2_X, σ^2_Y respectively. Let $\underline{X}_R = (X_{R_1},\ldots,X_{Rn_1})$ be the rounded sample where X_{Ri} is the rounded value of X_i corresponding to a rounding lattice with interval of width w_1 and lattice position c_1. $\underline{Y}_R = (Y_{R_1},\ldots,Y_{Rn_2})$ be the rounded sample where Y_{Ri} is the rounded value of Y_i corresponding to a rounding lattice with interval of width w_2 and lattice position c_2.

For testing the hypothesis $H_0:\mu_X = \mu_Y$, assuming $\sigma^2_X = \sigma^2_Y = \sigma^2$; the t-test statistic is given by:

$$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } S_p^2 = \frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1 + n_2 - 2} \qquad (3.3\text{-}13)$$

$\bar{X}$, $\bar{Y}$ are the sample means and S^2_X, S^2_Y, the usual estimates of the common variance σ^2.

In the first instance we first consider equal sample sizes, $n_1 = n_2 = n$, with both standard normal populations rounded according to the same rounding lattice (ie $w_1 = w_2 = w$, $c_1 = c_2 = c$). The test statistic for rounded data is given by:

$$t_R = \frac{\bar{X}_R - \bar{Y}_R}{\sqrt{\frac{S^2_{XR} + S^2_{XR}}{n}}} \qquad (3.3\text{-}14)$$

where

$$\bar{X}_R = \sum_i \frac{X_{Ri}}{n},\ \bar{Y}_R = \sum_i \frac{Y_{Ri}}{n},\ S^2_{XR} = \frac{1}{n-1}\sum_i (X_{Ri}-\bar{X}_R)^2,\ S^2_{YR} = \frac{1}{n-1}\sum_i (Y_{Ri}-\bar{Y}_R)^2$$

and both X_i and Y_i rounded to precision $r = w/\sigma$.

To obtain the approximate moments of t_R we use expansions for the monents of a two sample t–test statistic, for samples drawn from non–normal populations (Geary, 1947). Geary's results are general, allowing for different moments in the two populations, besides unequal sample sizes. Assuming equal sample sizes and the same degree of rounding for both populations we have

$$\begin{aligned} E[t_R] &= 0 \\ V[t_R] &= 1 + \frac{1}{n} + O(n^{-2}) \\ \sqrt{\beta_1}(t_R) &= 0 \\ \beta_2(t_R) &= 3 + \frac{3}{n} + O(n^{-2}) \end{aligned} \qquad (3.3\text{–}15)$$

To order n^{-1} the expressions (3.3–15) contain no population parameters.

If the normal populations are not subject to rounding we have

$$\begin{aligned} E[t] &= 0 \\ V[t] &= \frac{(n-1)}{(n-2)} = 1 + \frac{1}{n} + O(n^{-2}) \\ \sqrt{\beta_1}(t) &= 0 \\ \sqrt{\beta_2}(t) &= 3 + \frac{3}{n-3} = 3 + \frac{3}{n} + O(n^{-2}) \end{aligned} \qquad (3.3\text{–}16)$$

It follows from (3.3–15) and (3.3–16) that where both sample sizes are equal and rounded to the same rounding lattice, we expect the distribution of t to change very little for rounded data.

EXACT/SIMUL Results

As t_R is symmetrical about zero there is no difference between the upper and lower tail values of α_R. Nevertheless, values of α_R obtained from the SIMUL program will show slight variation due to sampling errors.

$n \leq 5$

Table (B.7) gives the range in values of α_R for n = 5, where r = 1.5 and 1.0.

For small values of n a major problem caused by rounding in the one sample t-test were the values $\pm\infty$ and 0/0. As in the one sample t-test, t_R = 0/0 was defined to be equal to zero. With the two sample t-test the probability of such values considerably reduced [Table (3.3.5)]. This together with the fact that t_R in the two sample case will be less discrete, than the one sample test, we would expect the two sample t-test to be more robust to rounding. The results for α_R confirmed this. For example, as shown by the results in Table (B.7) α and α_R were in close agreement for n = 5 where $r \leq 1.0$. For the one sample t-test this was only true for $r \leq 0.5$ [Table (B.1)]. Even for rounding as coarse as r = 1.5, the values of α_R and α were in reasonable agreement for the two sample t-test [Table (B.7)].

Table 3.3.5

Probability that $t_R = \pm\infty$ or 0/0 for samples of size n drawn from normal rounded populations in a two sample t-test

r	2.0		1.5		1.0	0.5	0.25
c / n	0	0.5	0	0.5	all c	all c	all c
2	0.260	0.209	0.153	0.147	0.073	0.019	0.005
3	0.106	0.047	0.033	0.027	0.007	0.005	0.000
4	0.047	0.011	0.009	0.005	0.001	0.000	
5	0.022	0.002	0.002	0.001	0.000		

n = 10 and 25

Table (B.8) gives the range in values of α_R for n = 10 and 25, where r = 2.0.

The robustness of this two sample test to very coarse rounding is striking. For $r \leq 2.0$ the values of α_R and α were in close agreement. As expected an increase in n from 10 to 25 caused an improvement in this agreement [Table (B.8)].

Table (3.3.6) gives the values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25.

Table 3.3.6

The values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25 in a two sample t-test.

	One tailed test						Two tailed test
α(%)	0.1/1.0/5.0		1.0/5.0		5.0		5.0
n	LT	UT	LT	UT	LT	UT	
5	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$
10	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$
25	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$	$r \leq 2.0$

The ranges of r given in Table (3.3.6) indicate how very robust the two sample t-test is, with respect to the level of significance for rounded data.

Unequal n's, c's and r's

The previous section was restricted to equal sample sizes with both populations rounded to the same rounding lattice. These are the conditions likely to be striven for in designing a comparison between means. However it is of interest to explore the situation when this restriction is not met.

Using Geary (1947) results, the approximate moments of t_R, where the rounding lattice is different in the two populations and the sample sizes are unequal indicated that:

(a) unequal sample sizes will have a negligible effect on the distribution of t_R

(b) different rounding lattices in the two population will cause the distribution of t_R to change very little from its form where both populations have been rounded the same.

From (a) and (b) above, we would expect the effect of rounding on the significance level of the test to be similar whether or not the two populations correspond to the same rounding lattice, and whether or not they have the same size. Simulation results for values of α_R, with differing rounding lattices and sample sizes confirmed this. For example, Table (3.3.7) shows the range in values of α_R where both sample sizes are equal to 5 and the two populations have been rounded to r = 1.5 and 0.1 respectively. These values of α_R are compared with the situation where both populations have been rounded to r = 1.5. As the table shows, the difference in the range of values of α_R, between the situation where both populations have precision equal to r = 1.5, or r = 1.5 and 0.1 respectively is small.

Table 3.3.7

Range of α_R values for sample sizes of size 5, where the two populations have been rounded to r = 1.5 and 0.1 respectively.

	Pop.		α(%) lower tail				α(%) upper tail			
	X	Y	0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
r	1.5	1.5	0.04-0.10	0.55-0.95	2.60-2.92	5.10-5.32	5.09-5.33	2.57-2.89	0.54-0.93	0.04-0.09
r	1.5	0.1	0.08-0.17	0.91-1.10	2.41-2.72	4.81-5.11	4.81-5.10	2.37-2.69	0.91-1.12	0.08-0.16

3.3.4 F-test for equality of two variances

Let $\underline{X} = (X_1,\ldots,X_{n_1})$ and $\underline{Y} = (Y_1,\ldots,Y_{n_2})$ be independent random samples of size n_1 and n_2 from normal populations X and Y, with means μ_X, μ_Y and variances σ^2_X, σ^2_Y respectively. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn_1})$ be the random sample where X_{Ri} is the rounded value of X_i corresponding to the rounding lattice with interval of width w_1 and lattice position c_1. $\underline{Y}_R = (Y_{R1},\ldots,Y_{Rn_2})$ be the rounded sample where Y_{Ri} is the rounded value of Y_i corresponding to a rounding lattice with interval of width w_2 and lattice position c_2.

For testing the hypothesis $H_0:\sigma^2_X = \sigma^2_Y = \sigma^2$, the F-test statistic is given by:

$$F = S^2_X/S^2_Y \qquad (3.3\text{-}17)$$

where S^2_X and S^2_Y are the usual estimates of the common variance σ^2.

We shall first consider equal sample sizes, $n_1 = n_2 = n$, with standard normal populations both rounded according to the same rounding lattice (ie $w_1 = w_2 = w$, $c_1 = c_2 = c$). The test statistic for rounded data is given by;

$$F_R = S^2_{XR}/S^2_{YR} \qquad (3.3\text{-}18)$$

where $S^2_{XR} = \frac{1}{n-1}\sum_i(X_{Ri}-\bar{X}_R)^2$, $S^2_{YR} = \frac{1}{n-1}\sum_i(Y_{Ri}-\bar{Y}_R)^2$ and both X_i and Y_i rounded to precision $r = w/\sigma$ under H_0.

We now use approximations to the first two moments of F_R (Gayen, 1950):

$$E[F_R] = 1 + \frac{1}{n}(\beta_{2R}-1) + O(n^{-2})$$

$$V[F_R] = \frac{1}{n}(\beta_{2R}-1) + O(n^{-2}) \qquad (3.3\text{-}19)$$

where β_{2R} is the measure of kurtosis for the two rounded normal populations.

Approximations to the first two moments of F_R, using (3.3–19) and Sheppard's corrections (3.3–4), are given by

$$E[F_R] = 1 + \frac{1}{n}\left[2 - \frac{\frac{r^4}{120}}{\left[1 + \frac{r^2}{12}\right]^2}\right] + O(n^{-2}) \qquad (3.3\text{-}20)$$

$$V[F_R] = \frac{2}{n}\left[2 - \frac{\frac{r^4}{120}}{\left[1 + \frac{r^2}{12}\right]^2}\right] + O(n^{-2})$$

If the normal populations are not subject to rounding we have:

$$E[F] = \frac{n-1}{n-3} = 1 + \frac{2}{n} + O(n)$$

$$V[F] = \frac{4(n-1)(n-2)}{(n-3)^2(n-5)} = \frac{4}{n} + O(n^{-2}) \qquad (3.3\text{-}21)$$

It follows from (3.3–20) and (3.3–21) that, where both sample sizes are equal and rounded to the same rounding lattice, we expect F_R to change very little for rounded data.

EXACT/SIMUL Results

Where both sample sizes are equal there is no difference between the upper and lower tail values of α_R. However, lower and upper tail values of α_R obtained from SIMUL program will show slight variation due to sample errors.

$n \leq 5$

Table (B.9) gives the range in values of α_R for n = 5, where r = 1.5, 1.0 and 0.5. Similar considerations apply to the distribution of F_R as to the distribution of t_R in section (3.3.3), namely that for small n the discontinuities in F_R will be numerous and that there is a possibility that either S^2_{XR} or S^2_{YR} is equal to zero. In particular, F_R will equal zero whenever S^2_{XR} is zero; it will equal infinity whenever S^2_{YR} is zero; it may be defined to be one when both S^2_{XR} and S^2_{YR} are zero. This definition of F_R = 0/0 equal to one seems sensible, as 0/0 indicates that the variances of X_R and Y_R are the same and under H_0 would expect F_R to equal one. The probability that F_R equals ∞ or 0/0 is identical to the probability that t_R equals $\pm\infty$ or 0/0; these probabilities are given in Table (3.3.1). Table (3.3.1) shows that F_R equal to $+\infty$ or 0/0 can occur with annoying frequency in very small samples, especially for coarse rounding (r > 1). For small values of n, F_R = 0 or $+\infty$ cause the tails of the distribution of F_R to contain greater probability than is expected in the unrounded situation. The effect of this is to inflate the α values. As expected, increasing n or by decreasing r will reduce the amount by which α values are inflated by F_R = 0 or $+\infty$. For example, as shown by the results for n = 5 in Table (B.9), the degree to which F_R = 0 or $+\infty$ inflates the value of α is dependent on the degree of precision r. At r = 2.0 all the four values of α are inflated in both tails, whereas at r = 1.0

only the value of $\alpha = 0.001$ has been inflated by $F_R = 0$ or ∞.

For small n, F_R will have a large number of discontinuities especially for coarse rounding. This will result in the distribution of F_R being a poor approximation to the distribution of F. We would expect then α and α_R values to be considerably different unless r is low. Our results for α_R confirmed this. For example, as shown by the results in Table (B.9), α and α_R are in reasonable agreement for $r \leq 0.5$ only.

n = 10 and 25

Tables (B.10) and (B.11) show the range in values of α_R for n = 10, r = 1.0, 0.5 and n = 25, r = 1.5, 1.0 respectively.

For these sample sizes, values of $F_R = \infty$ or 0/0 are now no real problem. For n = 10, $r \leq 0.5$ and n = 25, $r \leq 1.0$, the values of α_R and α are in close agreement. As expected an increase in n will improve this agreement. The similarity of the distributions of F and F_R for moderate size n was demonstrated by the close agreement between their means and variances. For example, for n = 25, the distributions of F_R and F had mean and variance 1.10, 0.225 and 1.10, 0.23 respectively. This confirms the results from the moments, which indicated that the mean and variance of F_R will be close to those of F for large n.

Table (3.3.8) gives the values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25.

Table 3.3.8

The values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25 in a F-test for equality of two variances.

	One tailed test						Two tailed test
α(%)	0.1/1.0/5.0		1.0/5.0		5.0		5.0
n	LT	UT	LT	UT	LT	UT	
5	$r \leq 0.5$	$r \leq 0.5$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$
10	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$
25	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$	$r \leq 1.5$

Unequal n's, c's and r's

The previous section was restricted to equal sample sizes and both populations rounded according to the same rounding lattice. These are the conditions likely to be striven for in designing a comparison between variances. Unequal sample sizes will have a very small influence on how the rounding process will affect F_R, unless n_1 and n_2 are very far apart. For this reason only different rounding lattices will be considered.

Let the normal populations X and Y be rounded according to rounding lattice with precision and lattice positions r_1, c_1 and r_2, c_2 respectively. Essentially X_R and Y_R will be two different populations. We shall assume that both samples are of size n. Expansions for the first two moments of $\bar{F}_R$, the test statistic where rounding is not the same in both populations, can be obtained from Gayen (1950).

$$E[\bar{F}_R] = \left[\frac{\sigma_R}{\sigma_R'}\right]^2 \left[1 + \frac{2}{n} + \left[\frac{\beta_{2R}'-3}{n}\right] + O(n^{-2})\right]$$

$$V[\bar{F}_R] = \left[\frac{\sigma_R}{\sigma_R'}\right]^4 \left[\frac{4}{n} + \frac{1}{n} + \{\beta_{2R}+\beta_{2R}'-6\} + O(n^{-2})\right] \qquad (3.3\text{-}22)$$

where X_R and Y_R have variance and kurtosis, σ^2_R, β_{2R} and $\sigma^{2'}_R$, β'_{2R} respectively.

Approximations to the first two moments of $\bar{F}_R$ using (3.3–22) and Sheppard's corrections (3.3–4) are given by:

$$E[\bar{F}_R] = \left[\frac{12+r_1^2}{12+r_2^2}\right]\left\{1 + \frac{1}{n}\left[2 - \frac{\frac{r_2^4}{120}}{\left[1 + \frac{r_2^2}{12}\right]^2}\right] + O(n^{-2})\right\}$$

$$V[\bar{F}_R] = \left[\frac{12+r_1^2}{12+r_2^2}\right]^2\left\{\frac{4}{n} - \frac{1}{n}\left[\frac{\frac{r_1^4}{120}}{\left[1 + \frac{r_1^2}{12}\right]} + \frac{\frac{r_2^4}{120}}{\left[1 + \frac{r_2^2}{12}\right]^2}\right] + O(n^{-2})\right\} \qquad (3.3\text{-}23)$$

Comparing (3.3–21) and (3.3–23) we have to the same order of approximation for $r \leq 2.0$

$$E[\bar{F}_R] \simeq \left[\frac{12+r_1^2}{12+r_2^2}\right] E[F] \ , \ V[\bar{F}_R] \simeq \left[\frac{12+r_1^2}{12+r_2^2}\right]^2 V[F] \qquad (3.2\text{-}24)$$

The first two moments of $\bar{F}_R$, indicate if $r_1 > r_2$, then rounding will cause the test statistic to have an increased mean and variance. For $r_1 < r_2$ there will be a decrease in the mean and variance. These moment results suggest that if

$r_1 \neq r_2$ then rounding will shift the distribution of F to the right or left, depending if $r_1 < r_2$ or $r_1 > r_2$. If $r_1 > r_2$ the significance level will be less than expected in the lower tail and greater than expected in the upper tail. For $r_1 < r_2$ the situation will be in reverse. Obviously this shift in distribution of F will be dependent on the values of r_1 and r_2. This effect will not diminish as n increases.

In order to gain insight into the behaviour of the significance level of the F-test for unequal rounding lattices, values of α_R were obtained for various r's and c's. To provide an indication of the possible effect, values of n and r given in Table (3.3.8) were considered. Table (3.3.8) gives the ranges of r that may be regarded as acceptable in the sense that α_R values fall within the ranges in section (3.3). Of interest is if these ranges of r are still suitable when the samples have not been rounded to the same rounding lattice.

For all three values of n, unequal r's resulted in a shift of the F distribution, as indicated by the moments of $\bar{F}_R$. As expected, this in general caused rounding to have greater effect than with equal r's. Table (3.3.9) shows a selection of results for n = 25.

Table 3.3.9

Values of α_R for F-test for samples of size 25, where values of the populations X and Y have been rounded to rounding lattices with c = 0 but different values of r.

X	Y	α(%) lower tail				α(%) upper tail			
r_1	r_2	0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
1.5	1.5	0.12	1.08	2.58	5.06	5.14	2.60	1.08	0.14
1.5	1.0	0.07	0.63	1.58	3.15	7.70	3.95	1.70	0.18
1.5	0.25	0.04	0.42	1.06	2.19	10.53	5.81	2.58	0.31
0.25	1.5	0.31	2.60	5.82	10.50	2.18	1.05	0.44	0.04

The results in Table (3.3.9) demonstrate how different r values will shift the F distribution to the right or left depending if $r_1 < r_2$ or $r_1 > r_2$. As the difference in r_1 and r_2 increased so did the shift in the distribution, causing rounding to have a greater effect on the α values. Although for $r_1 = r_2 = 1.5$ the values of α and α_R are in reasonable agreement, this is not always so for $r_1 \neq r_2$. The results in Table (3.3.9) are typical of the values of α_R found for $r_1 \neq r_2$.

The simulated results for the values of α_R, suggested a 'rule of thumb' that may be applied when the rounding is not the same in both populations:

Let $$R = \frac{12 + r_1^2}{12 + r_2^2} \qquad \text{where } r_1 \geq r_2$$

The ranges of r given in Table (3.3.8) are only suitable for r_1 and r_2 for n = 5 if $R \leq 1.08$. For n = 10 and 25 the corresponding values are $R \leq 1.06$ and

$R < 1.04$ respectively.

Different lattice positions ($c_1 \neq c_2$) were found to have far less effect than $r_1 \neq r_2$.

3.3.5 Analysis of Variance (ANOVA)

The consequences when the assumptions for the ANOVA are not satisfied have been studied by many authors. However to date the only research into the effect of rounding on the ANOVA has been by Riley, Bekele and Shrewsbury (1983), who considered only specific examples; they made no attempt to obtain general conclusions. In this section, unlike in the previous authors, a simulation method is used to investigate the sensitivity of the significance level to rounded data in the one and two-way ANOVA.

One-way Analysis of Variance - Fixed Effects Model

The structure we shall assume for the one-way layout fixed effect model is given by:

$$X_{ij} = \mu + \alpha_i + e_{ij} \qquad i = 1,\ldots,k \qquad j = 1,\ldots,n$$

where μ is the overall mean

α_i is the ith sample effect and are fixed constants

e_{ij} are errors distributed normally and independently about zero with the same variance σ^2

This study was restricted to equal sample sizes. For testing the hypothesis $H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_k$, the test statistic is given by:

$$F = \frac{Q_1/(k-1)}{Q_2/(nk-k)} \qquad (3.3\text{-}25)$$

where $Q_1 = \sum_{i=1}^{k} n(\bar{X}_{i.} - \bar{X}_{..})^2$ the between sum of squares

$Q_2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_{i.})^2$ the within sum of squares

Under H_0 F(3.3–25) is distributed as a central F distribution with k–1 and nk–k degrees of freedom.

Let the ith sample be drawn from the ith normal population which has been rounded according to a rounding lattice with rounding interval w_i and lattice position c_i. We shall first assume that all normal populations have been rounded to the same rounding lattice (ie $w_1 = w_2 = \ldots w_R = w$; $c_1 = c_2 = \ldots c_k = c$). The test statistic for rounded data is given by

$$F_R = \frac{Q_{1R}/(k-1)}{Q_{2R}/(nk-k)} \qquad (3.3\text{-}26)$$

where $Q_{1R} = \sum_{i=1}^{k} n(\bar{X}_{Ri.} - \bar{X}_{R..})^2$, $Q_{2R} = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{Rij} - \bar{X}_{Ri.})^2$ and

X_{ij} rounded to precision $r = w/\sigma$ and lattice position c.

Once again for convenience the normal populations will be assumed to have zero means and variances equal to one. Again using results of Gayen (1950) we have

$$E[F_R] = 1 + \frac{2}{N} + O(N^{-2})$$
$$V[F_R] = \frac{2}{k-1} + \frac{2}{N(k-1)}\,[5+k-(\beta_{2R}-3)] + O(N^{-2}) \qquad (3.3\text{-}27)$$

where $N = nk$ and β_{2R} is the measure of kurtosis for the rounded normal populations.

The approximations to the first two moments of F_R using (3.3–27) and Sheppard's corrections (3.3–4) are given by:

$$E[F_R] = 1 + \frac{2}{N} + O(N^{-2})$$
$$V[F_R] = \frac{2}{k-1} + \frac{2}{N(k-1)}\left[5 + k + \frac{\frac{r^4}{120}}{\left[1 + \frac{r^2}{12}\right]^2}\right] + O(N^{-2}) \qquad (3.3\text{-}28)$$

If the normal populations are not subject to rounding we have

$$E[F] = \frac{N-k}{N-k-2} = 1 + \frac{2}{N} + O(N^{-2})$$
$$V[F] = \frac{2(N-k)^2(N-3)}{(k-1)(N-k-2)^2(N-k-4)} = \frac{2}{k-1} + \frac{2}{N(k-1)}\,(5+k) + O(N^{-2}) \qquad (3.3\text{-}29)$$

Comparison of (3.3–28) with (3.3–29) implies that when N is large we expect the distribution of F to change very little under rounding.

SIMUL Results

For ANOVA only the SIMUL program was used to obtain results. Values of α_R were obtained for k = 3, 5 and 10, where n = 5, 10 and 25.

k = 3

Table (B.12) shows the range in values of α_R for k = 3, where n = 5, 10 and 25.

For this value of k the discontinuities in F_R can be numerous and there is a possibility that either Q_{1R} or Q_{2R} are equal to zero. In particular, F_R will equal zero whenever Q_{1R} is zero; it will equal infinity whenever Q_{2R} is zero; it may be defined to be one when both Q_{1R} and Q_{2R} are zero. The last mentioned situation cannot occur unless all the values in the (k×n) data set are identical and this will have a very small probability of occurrence. For coarse rounding ($r > 1$) F_R equal to zero or infinity can occur with annoying frequency for small n. The result is that values of F_R equal to zero or infinity can distort the α values. As the probability of F_R equal to zero is greater than F_R equal to infinity, the lower tail values of α will be more affected by the rounding process. As shown by the results in Table (B.12) the degree to which $F_R = 0$ or ∞ distorts the values of α is dependent on n and r. For n = 5, the lower tail values of α are severely affected and to some extent the upper tail values for $r \geq 1.0$ while for n = 10, it is only the lower tail values of α affected for $r \geq 1.5$.

The approximations for the first two moments of F_R indicate that the moments will change very little for rounded data, where N is large. As shown by the

results for the moments of F_R given in Table (B.12), this was true for N as small as 15. However, rounding will cause the distribution of F_R to become discrete. It was the discrete nature of F_R that made the α values distorted. Generally the lower tail values of α were more affected by this discretization of the F distribution. Although we are interested in the upper tail values of α, as we are dealing with a one-tailed test, the lower tail values will show an indication of how the F distribution behaves with respect to rounding.

k = 5 and 10

Table (B.13) shows the range in values of α_R for k = 5, 10 where n = 5.

For these larger values of k, there will be a corresponding increase in N (nk). This larger value of N will mean that the F_R distribution will have less discontinuities and a lower probability that F_R = 0/0, ∞ and 0. The results in Table (B.13) show how this larger value of k generally improve the agreement between the α and α_R values. This being more so in the lower tail.

To obtain a more accurate recommendation for when the degree of precision r may be regarded as acceptable, values of α_R were obtained for k = 3 and 4, for n ranging between 6 and 10. Table (3.3.10) gives the values of the degree of precision r that were found acceptable.

Table 3.3.10

The values of the degree of precision r that may be regarded as acceptable for N = nk in a one and two way analysis of variance.

α(%)	0.1/1.0/5.0	1.0/5.0	5.0
N* = 15	r ≤ 1.5	r ≤ 1.5	r ≤ 1.5
N ≥ 16	r ≤ 2.0	r ≤ 2.0	r ≤ 2.0

*N = nk when k = 3,4,... and n = 5,6,...

The ranges in Table (3.3.10) apply only to the upper tail values of α_R, as we are dealing with a one-tailed test. However the lower tail values of α are far more distorted by rounding and the ranges of r given in the table do not necessarily apply.

Unequal r's and c's

Consider the situation where the k samples of size n have been drawn from normal populations rounded according to different rounding lattices. The ith sample will have rounding precision r_i and lattice position c_i. Essentially the samples will have been drawn from different populations with parameters μ_{Ri}, σ^2_{Ri}, $\sqrt{\beta_1}_{Ri}$, β_{2Ri} (i=1,...,k). As indicated by the results in Chapter 2 for normal populations with r ≤ 2.0 the discrepancy between these four parameters and their values for unrounded data is not very serious. As a result we would expect different rounding in the k populations to have a similar effect on the α's as the situation when the populations have all been rounded the same. For example

consider the variances. With different rounding in the populations the variances σ^2_{Ri} are approximately equal to $1 + r^2/12$. From Box (1954) for equal sample sizes, the maximum effect on the significance level of the test will be when the variances σ^2_{iR} are in the ratio $1 : 1 : 1 : \ldots : 1 : \theta$, where θ is the $\max(\sigma^2_{iR})/\min(\sigma^2_{iR})$. For $r = 2.0$, the maximum value of θ will occur at $r_1 = 2.0$, $r_2 = 0$, ..., $r_k = 0$, giving $\theta = 1.33$. Thus at the most severe rounding considered the maximum value of θ is only 1.33. With such a small value of θ, Box's results indicate that unequal variances will have little effect on the significance level of the test.

Simulation results for values of α_R, with differing rounding lattices in the populations were obtained. Table (3.3.11) shows a selection of results for the upper tail values of α_R for populations rounded so that θ has its maximum value. These are compared with values of α_R where the rounding is the same for all populations (ie $\theta = 1$). The results in the table are typical of the values of α_R found for unequal rounding lattices.

Table 3.3.11

Values of α_R where the populations have been rounded according to different rounding lattices.

Anova k×n	r	θ	Upper tail α(%)		
			5.0	1.0	0.1
3×5	1.5	1*	4.7	0.9	0.09
		1.19†	5.1	0.1	0.10
3×10	2.0	1	5.2	1.1	0.09
		1.33†	5.0	1.0	0.10
3×25	2.0	1	5.0	1.0	0.09
		1.33	5.0	1.0	0.11
5×5	2.0	1	4.4	0.7	0.10
		1.33	5.0	1.0	0.10
5×10	2.0	1	4.8	1.0	0.09
		1.33	5.0	1.0	0.10
5×25	2.0	1	5.0	1.0	0.11
		1.33	5.0	1.0	0.12

* $\theta = 1$ – all samples rounded to same r

† maximum value of θ for given r.

As Table (3.3.11) shows, the only noticeable effect of unequal precision in the populations is for small N (ie k = 3, n = 5). As the value of N increases the effect of unequal r's diminishes.

In general the simulation results for α_R indicated that different rounding in the k populations will have no more effect on the significance level of the test than when the populations have the same rounding.

Two-way Analysis of Variance - Fixed Effects Model

The structure we assume for the two-way layout fixed effects model is given by:

$$X_{ij} = \mu + \alpha_i + \beta_i + e_{ij} \qquad \begin{matrix} i = 1,\ldots,k \\ j = 1,\ldots,n \end{matrix}$$

where μ is the overall mean

α_i is a fixed effect due to ith row

β_j is a fixed effect due to jth column

e_{ij} are errors distributed normally and independently about zero with the same variance σ^2

For testing the hypothesis $H_0:\alpha_1 = ,\ldots, \alpha_k$ the test statistic is given by

$$F = \frac{Q_1/(k-1)}{Q_E/(n-1)(k-1)} \qquad (3.3\text{-}30)$$

where $$Q_1 = n \sum_{i=1}^{k} (\bar{X}_{i.}-\bar{X}_{..})^2, \quad Q_E = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..})^2$$

Under H_0 F is distributed as a central F distribution with k-1 and (n-1)(k-1) degrees of freedom.

Let X_{ij} be rounded to the same rounding lattice for all i,j where the rounding interval is w and lattice position c. The test statistic for rounded data is given by:

$$F_R = \frac{Q_{1R}/(k-1)}{Q_{ER}/(n-1)(k-1)} \qquad (3.3\text{-}31)$$

where $Q_{1R} = n \sum_{i=1}^{k} (\bar{X}_{Ri.} - \bar{X}_{R..})^2$, $Q_{ER} = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{Rij} - \bar{X}_{Ri.} - \bar{X}_{R.j} + \bar{X}_{R..})^2$

and rounding precision $r = w/\sigma$ and lattice position c.

The difference between the values of F for the one-way (3.3-25) and two-way ANOVAs is caused by the denominators Q_2 and Q_E. Hence the only difference in the numerical calculations performed between the one and two way ANOVAs is the smaller residual term in the two-way test (ie a smaller denominator in the F ratio). As shown by the one way ANOVA results, the ratio of quadratic forms used is fairly insensitive to rounding.

We would expect the situation to be similar for the two way ANOVA. As Q_E is a more 'complicated' quadratic form than Q_2, the discontinuities of F_R in the two way ANOVA will be less numerous than for the one way situation. This will result in the two way ANOVA being less effected by rounding, especially for small N (nk).

Values of α_R were obtained for k = 3, 5 and 10. Table (B.14) compares the values of α_R obtained for k = 3 and n = 5 in the one and two way ANOVA. As expected the results were very similar for both tests. Although only the results for k = 3 and n = 5 are shown, they do faithfully represent the entire body of results. In general the values of α_R were very similar for both one and two way ANOVA. Closer investigation of the simulation results indicated that in the two way ANOVA the significance level was slightly less affected by rounding. This was

noticeable for small values of N (nk) the ranges of r given in Table (3.3.10) were also found to be suitable for the two-way ANOVA.

3.4 Discussion and Conclusions

This chapter has investigated the effect of rounding on the significance level of a statistical test. The results have been given in such a way that the reader can readily determine the general trend of what happens to the significance level as the degree of precision in the data is allowed to vary. Five basic tests were investigated. The behaviour of one and two-way ANOVAs has provided insight into the effect of rounding on this general statistical procedure.

The t and F tests were as robust to rounding as each other. In both tests rounding as coarse as r = 1.0 and 1.5 in samples of size 10 and 25 respectively were found to give suitable levels of significance. However, different rounding lattices in the samples made the F-test more sensitive to rounding.

Of all the tests considered the two sample t-test was the most insensitive to rounding. For all values of n considered, rounding as coarse as r = 1.5 gave acceptable levels of significance. Different rounding lattices in the samples did not alter the effect of rounding to any extent.

The chi-squared test was the most sensitive to rounded data, the main reason being that rounding increased the mean and variance of the test statistic. This correspondingly caused the significance levels to be distorted. Satisfactory levels of significance were obtainable only for low values of r (ie r = 0.25 or 0.5).

The section dealing with the ANOVA demonstrated how this statistical technique is robust to rounding. The results for both the one and two way ANOVAs showed how data could be rounded appreciably without any serious change in the significance level of the test. The ratio of quadratic forms which made up the F statistic were found to be extremely robust to rounding. The quadratic forms used in the one and two-way layouts are similar to the quadratic forms in higher way layouts. Our results suggest that generally the ANOVA technique is insensitive to rounded data with respect to the level of significance.

In this chapter ranges of r that give satisfactory levels of significance have been given for three values of n, ie n = 5, 10 and 25. In order to present a more detailed picture of the effect of rounding on the significance level of a test, values of α_R were obtained for other values of n. Table (3.3.12) shows the ranges of r and n which may be regarded as acceptable. To keep the computing to within manageable bounds, only values of α = 0.05 and 0.01 were considered. It is clear from the results in Table (3.3.12) that all tests except the chi-squared are robust to rounding.

As mentioned in the literature review, various rules have been suggested for the degree of precision that should be used when recording data. It is of interest to see how suitable these rules are. Rules for rounding have been given by several authors. There seems to be no standard set of rules. The most commonly quoted is that r should not exceed $\frac{1}{4}$ (eg Eisenharht, 1947) or the less stricter rule that r should not exceed $\frac{1}{2}$ (eg Nicholson, 1979). Although the results in Table (3.3.12) show that for all tests except the chi-squared, using r less than $\frac{1}{2}$ or $\frac{1}{4}$ will give satisfactory levels of significance, they are generally too conservative. An acceptable level of significance is possible with far more coarse rounding.

There are many persons, both sophisticated statisticians and others who would like to know when they could apply standard normal theory tests to rounded data. The results of this study indicate the extent to which data may be rounded without adversely affecting the level of significance of the test. In most situations we can use far less precision in rounding than originally realised and still apply standard tests.

Table 3.3.12

Values of (n,r) which may be regarded as acceptable for five standard tests

α(%)	5.0/1.0		5.0		two tailed
test	lower tail	upper tail	lower tail	upper tail	5.0
		(=5) r≤0.25			(=5) r≤0.5
χ^2	(5,50) r≤0.5	(6,24) r≤0.5	(5,50) r≤0.5	(5,25) r≤0.5	(6,9) r≤1.0
	(≥51) r≤0.25	(≥25) r≤0.25	(≥51) r≤0.25	(≥26) r≤0.25	(≥10) r≤0.5
one	(5,7) r≤0.5	(5,7) r≤0.5			(5,8) r≤0.5
sample	(8,13) r≤1.0	(8,13) r≤1.0	same as	5.0/1.0	(9,13) r≤1.0
t	(14,29) r≤1.5	(14,29) r≤1.5			(14,29) r≤1.5
	(≥30) r≤2.0	(≥30) r≤2.0			(≥30) r≤2.0
two	(5,6) r≤1.5	(5,6) r≤1.5	(≥5) r≤2.0	(≥5) r≤2.0	(≥5) r≤2.0
sample	(≥7) r≤2.0	(≥7) r≤2.0			
t					
	(5,10) r≤1.0	(5,10) r≤1.0	(5,9) r≤1.0	(5,9) r≤1.0	(5,9) r≤1.0
F	(≥11) r≤1.5	(≥11) r≤1.0	(≥10) r≤1.5	(≥10) r≤1.5	(≥10) r≤1.5
Anova	-	N=15 r≤1.5	-	N≤15 r≤1.5	-
		N≥16 r≤2.0		N≥16 r≤2.0	

Note: notation: (a,b) → $a \leq n \leq b$

(=a) → $n = a$

(≥a) → $n \geq a$

* $N = kn$ where $k = 3,4,\ldots$

$n = 5,6,\ldots$

CHAPTER 4

THE EFFECT OF ROUNDING ON THE POWER LEVEL OF CERTAIN NORMAL TEST STATISTICS

4.1 Introduction

The effect of rounding on the significance level of tests has been studied in Chapter 3. The results indicate that in most situations, the significance levels of these tests are insensitive to rounding. For each test a range of r was recommended that gave acceptable levels of significance under rounding. This chapter investigates whether the powers of these tests are adversely affected for those recommended values of r given in Chapter 3. At the end of the chapter a method for compensating for the effect of rounding in the chi–squared test is discussed.

In the literature many investigations have studied the effect that departure from normality has on the power of standard statistical tests. However to date no work has looked at the effect of reduced precision on the power of a test for rounded normal data.

4.2 Description of the Investigation

The power was evaluated for each statistical test in Chapter 3. To keep the investigation within reasonable bounds the power of the tests under rounding was found for $\alpha = 0.05$ for one tailed tests. However for some tests the power was also found for $\alpha = 0.01$ and 0.001. The power was evaluated mainly for values of r for which the significance level of the test was found acceptable, for n = 5, 10 and 25. The usual lattice positions c = −0.5, ..., 0.5 were considered for each value of r. The value of $\alpha = 0.05$ was chosen as this is normally the first level of significance at which the null hypothesis is rejected. To consider in detail the power for other values of α would have required much computer time. The range

of n considered closely reflects the sample sizes most commonly used in practice. The sample size of 25 was found in most situations to give a good indication of the effect of rounding on the power for large samples. For the two sample t-test, F-test for equality of variances and F-test in one and two way analysis of variance, only samples of equal size will be considered. The power of each test under rounding was evaluated for values of the alternative hypothesis H_1, corresponding to powers of 0.3, 0.5, 0.7 and 0.95 under normal theory conditions.

In the previous chapter the significance level of the test for rounded data was examined by determining: (1) approximations to the sampling moments of the test statistics; (2) the exact distribution of the test statistic for rounded data; (3) an estimate of the sampling distribution of the test statistic for rounded data by Monte Carlo Methods. However, for most test statistics it was found not to be necessary to obtain approximations to the sampling moments, the reason being that the effects of rounding on the test statistics under H_0 and H_1 were similar.

Two FORTRAN programs were written for the analysis. The program PEXACT generated every possible sample of size n from a normal population that has been rounded according to a rounding lattice with rounding interval w and lattice position c. The required test statistic was calculated and the power of the test for rounded data was obtained under normal theory conditions. The program PSIMUL generated N random samples of size n from a normal population rounded to a specific w and c. As in the program PEXACT the power of the test for rounded data was obtained. Both programs also gave the mean and variance of the test statistic for rounded data.

The PEXACT program required an exorbitant amount of computer time. As a result it was decided to use the PSIMUL program to generate the required powers. The PEXACT program was used only for checking the results of PSIMUL. The results from the simulation were based on 100,000 iterations, ie 100,000 values of each test statistic were generated for estimating each power under rounding. Of course, the results obtained for the powers from the simulation are subject to sampling errors. For simulations of 100,000 these will be small. The standard errors for the estimates of powers equal to 0.3, 0.5, 0.7 and 0.95 will be all less than $1.60(10)^{-3}$.

Quality of Results

Both the PEXACT and PSIMUL programs were tested to check the validity of their results. For example an independent check on the results given by PSIMUL was provided by obtaining the powers for each test statistic when the normal populations were not subject to rounding. They were found to agree very closely with the expected results. Another check was established by comparing a selection of results from both PEXACT and PSIMUL programs.

4.3 Test Statistics

In this section it is assumed that; when the null hypothesis (H_0) is satisfied, the normal distributions have mean zero and variance one. The non-null situation (H_1) was handled by adjusting the required parameters in the standard normal distribution to give the required power under normal theory conditions. Throughout this section P will denote the power of the test for samples from unrounded normal populations, while P_R will be the resulting power of the test for

samples from rounded normal populations. P_R may be regarded as the true power of the test when the data have been rounded. The results are presented as follows. For each test statistic:

(i) Before discussing the simulation results, the behaviour of the test statistic for rounded data under H_1 will be examined.

(ii) The power values P_R for which the significance level of the test was found acceptable under rounding, for n = 5, 10 and 25 are discussed. For convenience the P_R results are tabulated.

Appendix B gives a list of all the output produced by the PSIMUL program.

4.3.1 One sample t-test

Let $\underline{X} = (X_1,\ldots,X_n)$ be a random sample of size n from a normal population X. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the value of X_i corresponding to a rounding lattice with interval of width w and lattice position c. For testing the hypothesis $H_0:\mu = \mu_0$ vs $H_1:\mu \neq \mu_0$ the test statistic is given by (3.3-1). Under H_1

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}(\delta) \qquad (4.3\text{-}1)$$

where $t_{n-1}(\delta)$ is a non-central t distribution with non-centrality parameter

$\delta = \frac{\sqrt{n}}{\sigma}(\mu-\mu_0)$ and μ is the value of the population mean under H_1

and under rounding

$$t_R = \frac{\bar{X}_R - \mu_0}{S_R/\sqrt{n}} \qquad (4.3\text{-}2)$$

where $\bar{X}_R$ and S_R are defined as in (3.3–2).

Under H_0, for values of r within the recommended ranges given in Table (3.3.2), the distribution of t_R (3.3–2) closely approximated that of a t distribution. Under H_1 we would expect a similar situation, namely that the distribution of t_R (4.3–2) will be in close agreement with a non–central t distribution for these recommended values of r. As the population is subject to rounding the non–centrality parameter will be:

$$\delta_R \simeq \frac{\sqrt{n}(\mu_R - \mu_0)}{\sigma_R} \qquad (4.3\text{-}3)$$

As the population is normal, good approximation to μ_R and σ_R are given by Sheppard's corrections. Applying Sheppard's corrections to approximate μ_R and σ_R we have

$$\delta_R \simeq \frac{\sqrt{n}(\mu - \mu_0)}{\sigma\sqrt{1 + \frac{r^2}{12}}} = \frac{\delta}{\sqrt{1 + \frac{r^2}{12}}} < \delta \qquad (4.3\text{-}4)$$

From (4.3–4) we would expect the distribution of t_R to be in close agreement with a non–central t distribution, $t(\delta_R)$, where δ_R is less than δ. This reduction in the non–centrality parameter caused by rounding will result in the test becoming less powerful for rounded data. If values of r satisfy the recommendations given in Table (3.3.2), then it is reasonable to approximate the distribution of t_R by $t(\delta_R)$.

This can be illustrated as follows.

For various values of r within the recommended ranges given in Table (3.3.2) for $\alpha = 0.05$ (one tailed)

(i) the mean and variance of the distribution of t_R were obtained by simulation and compared with those for a $t(\delta_R)$ distribution.

(ii) the powers of the test for values of H_1 corresponding to powers of 0.3, 0.5, 0.7 and 0.95 under normal theory conditions were obtained by simulation and compared with those given by the distribution of $t(\delta_R)$.

Table 4.3.1

Range of values of P_R at $\alpha = 0.05$, mean and variance of t_R when n = 10 and r = 1.0 for a one sample t-test

Power	P_R (simulation)		$P_R(t(\delta_R))$	$E[t_R]^*$		$V[t_R]^*$	
P	lower tail	upper tail	lower & upper tail	simulation†	$t(\delta_R)$	simulation†	$t(\delta_R)$
0.30	0.279-0.299	0.276-0.298	0.285	1.27	1.28	1.40	1.39
0.50	0.474-0.489	0.464-0.489	0.474	1.87	1.87	1.52	1.55
0.70	0.663-0.684	0.663-0.683	0.669	2.47	2.47	1.78	1.74
0.95	0.933-0.940	0.935-0.941	0.931	3.75	3.76	2.38	2.32

* The mean and variance are given only for the upper tail. The values for lower tail were very similar with a change in sign for $E[t_R]$.

† The simulated values of the mean and variance differ only in their third decimal place for values of c.

Table 4.3.2

Range of values of P_R at $\alpha = 0.05$, mean and variance of t_R when n = 25 and r = 1.5 for a one sample t-test

Power	P_R (simulation)		$P_R(t(\delta_R))$	$E[t_R]^*$		$V[t_R]^*$	
P	lower tail	upper tail	lower & upper tail	simul-ation	$t(\delta_R)$	simul-lation	$t(\delta_R)$
0.30	0.259-0.281	0.257-0.284	0.265	1.09-1.10	1.09	1.12-1.13	1.12
0.50.	0.435-0.466	0.443-0.464	0.446	1.60-1.61	1.61	1.14-1.15	1.13
0.70	0.624-0.654	0.623-0.651	0.635	2.10-2.11	2.12	1.19-1.20	1.18
0.95	0.911-0.924	0.910-0.921	0.915	3.20-3.21	3.21	1.32-1.33	1.33

* The mean and variance are given only for the upper tail. The values for lower tail were very similar with a change of sign for $E[t_R]$.

Tables (4.3.1) and (4.3.2) show a selection of results for $\alpha = 0.05$. They illustrate the close agreement between the simulated results and those obtained from the distribution of $t(\delta_R)$. The value of the means and variances indicate that the distribution of $t(\delta_R)$ will closely approximate that of t_R. This is also evident by the close agreement of the P_R values from simulation and the distribution of $t(\delta_R)$. To obtain the P_R values from the distribution of $t(\delta_R)$, tables given by Owen (1965) were used.

The range in the simulated values of the mean, variance and P_R values is caused by the lattice position c.

Some values of P_R were also obtained for $\alpha = 0.01$ and 0.001. Close agreement between the distributions of $t(\delta_R)$ and t_R was also found for these two levels of

significance, if the value of r was in the recommended range. It is reasonable to conclude that if the value of r is in the recommended range given in Table (3.3.2) the distribution of $t(\delta_R)$ will be a good approximation to t_R. For values of r outside the range of values given in Table (3.3.2) the simulation results showed that the distribution of t_R became progressively unlike a non-central t distribution.

Simulation Results

In section (3.3.1) the recommended ranges of r for which the significance level of the test was found acceptable under rounding for $\alpha = 0.05$ (one tailed) were: $r \leq 0.5$ when $n = 5$, $r \leq 1.0$ when $n = 10$ and $r \leq 1.5$ when $n = 25$. The power of the test will be generally more affected by the rounding process at the maximum value of r allowed. As a result P_R values for the maximum value of r are of most interest. This would indicate the worst possible effect that rounding can have on the power of a test within the recommended range of r. For $\alpha = 0.05$, values of P_R for $n = 5$ at $r = 0.5$, $n = 10$ at $r = 1.0$ and $n = 25$ at $r = 1.5$ are shown in Tables (4.3.1 to 4.3.3).

Table 4.3.3

Range of values of P_R for a single sample t-test for $\alpha = 0.05$, $n = 5$, $r = 0.5$

P	P_R	
	lower tail	upper tail
0.30	0.294-0.306	0.292-0.304
0.50	0.494-0.505	0.491-0.502
0.70	0.693-0.701	0.692-0.699
0.95	0.946-0.949	0.947-0.949

As expected, rounding has caused a reduction in the power of the test. As indicated by (4.3-4), as the value of r increased, the reduction in the power becomes greater. In general the results indicate that the four power levels are not adversely affected by rounding if the values of r are in the recommended ranges for $\alpha = 0.05$ given in Table (3.3.2).

Although in this section the power of t_R test has been found by simulation, an estimate of this power can be obtained from tables or from an approximation for values of r. Assuming that values of r are within the ranges given in Table (3.3.1) then tables or a suitable approximation can be used to estimate the power P_R, which is given by $P[t(\delta_R) > t_{n-1,\alpha}]$. For example, tables given by Owen (1965) or an approximation to the cumulative distribution of the non-central t in Johnson and Kotz (1970) may be used.

4.3.2 Chi-squared test for a variance

Let $\underline{X} = (X_1,\ldots,X_n)$ be a random sample of size n from a normal population X. Let $\underline{X}_R = (X_{Ri},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the value of X_i corresponding to a rounding lattice with interval of width w and lattice position c. For testing $H_0:\sigma^2 = \sigma_0^2$ vs $H_1:\sigma^2 \neq \sigma_0^2$ the test statistic is given by (3.3-7). Under H_1

$$\chi^2 = \frac{(n-1)S^2}{\sigma_1^2} \sim \chi^2_{n-1} \qquad (4.3\text{-}5)$$

where χ^2_{n-1} is a chi-squared distribution with n-1 degrees of freedom and σ_1^2 is the value of σ^2 under H_1.

Under rounding

$$\chi^2_R = \frac{(n-1)S^2_R}{\sigma_1^2} \qquad (4.3\text{-}6)$$

where S^2_R is defined as in (3.3-8).

For both H_0 and H_1 the test statistics are essentially the same, each having a chi-squared distribution with n-1 degrees of freedom. Hence rounding will have the same effect on the test statistic under H_0 as for H_1. From section (3.3.2), we know the main effect of rounding will be to cause the distribution of the test statistic (4.3-5) to shift to the right. As r increases so does the size of the shift.

For any given sample of rounded data, w and c are fixed. For $\sigma \in H_0$ the degree of precision is $r = w/\sigma_0$, while for $\sigma \in H_1$ it is $r_1 = w/\sigma_1$. Thus giving a relationship between r and r_1 of the form $r_1 = r\sigma_0/\sigma_1$.

If (i) $H_1:\sigma^2 = \sigma_1^2 > \sigma_0^2$ then $r_1 < r$

(ii) $H_1:\sigma^2 = \sigma_1^2 < \sigma_0^2$ then $r_1 > r$

Unlike hypothesis tests for the mean μ, those concerned with the variance σ^2, will not have a constant degree of precision. The degree of precision will be dependent on the 'true' value of the parameter σ. For upper tailed tests (i), rounding will have less effect under H_1 than under H_0. In lower tailed tests (ii), the situation will be in reverse. The shift in the distribution to the right caused by rounding will reduce the power in the lower tail and increase it in the upper tail.

Simulation Results

In section (3.3.2) the recommended range of r for which the significance level of the test was found acceptable under rounding for α = 0.05 (one tailed) was $r \leq 0.5$ for n = 5, 10 and 25. As in the t-test, the P_R values for the maximum value of r are of most interest. For α = 0.05 values of P_R obtained by simulation for n = 5, 10 and 25, at r = 0.5 are shown in Tables (4.3.4 to 4.3.6).

Table 4.3.4

Range of values of P_R for a chi-squared test for α = 0.05, n = 5 and r = 0.5

P	P_R	
	lower tail	upper tail
0.30	0.306-0.308	0.307-0.308
0.50	0.496-0.502	0.506-0.507
0.70	0.662-0.684	0.704-0.705
0.95	0.889-0.923	0.950-0.950
α = 0.05	0.058-0.058	0.055-0.055

Table 4.3.5

Range of values for P_R for a chi-squared test for $\alpha = 0.05$, n = 10 and r = 0.5

P	P_R	
	lower tail	upper tail
0.30	0.272-0.275	0.310-0.312
0.50	0.457-0.458	0.508-0.509
0.70	0.647-0.649	0.704-0.705
0.95	0.917-0.919	0.951-0.952
$\alpha = 0.05$	0.046-0.046	0.056-0.056

Table 4.3.6

Range of values of P_R for a chi-squared test for $\alpha = 0.05$, n = 25 and r = 0.5

P	P_R	
	lower tail	upper tail
0.30	0.266-0.268	0.318-0.319
0.50	0.452-0.454	0.517-0.519
0.70	0.647-0.650	0.711-0.713
0.95	0.922-0.925	0.951-0.952
$\alpha = 0.05$	0.041-0.045	0.059-0.061

The P_R values in Tables (4.3.4 to 4.3.6) illustrate how rounding has caused the power to be reduced in the lower tail and increased in the upper tail. As expected the difference between the P and P_R values is far greater in the lower than in the upper. Although in upper tail tests rounding can cause the test to be more powerful, there is also a corresponding increase in the significance. In general the results indicate that the four power levels are not adversely affected by

rounding if values of r are in the recommended ranges for $\alpha = 0.05$ (one tailed) given in Table (3.3.4).

For values of r outside the range of values given in Table (3.3.4) for $\alpha = 0.05$, the simulation results showed how the change in power can be severe, especially in the lower tail. This is illustrated by the values of P_R given in Table (4.3.7) for $n = 10$, where $r = 1.0$.

Table 4.3.7

Range of values of P_R for a chi-squared test for $\alpha = 0.05$, n = 10 and r = 1.0

P	P_R	
	lower tail	upper tail
0.30	0.191-0.196	0.323-0.325
0.50	0.325-0.339	0.515-0.517
0.70	0.477-0.502	0.705-0.707
0.95	0.761-0.803	0.950-0.951
$\alpha = 0.05$	0.033-0.033	0.067-0.068

In this section P_R values have been simulated for $\alpha = 0.05$. However, the behaviour of the test statistic under rounding is the same for all α. Thus we would expect the change in power caused by rounding to be similar for $\alpha = 0.01$ and 0.001. Furthermore, if the values of r are restricted to the ranges given in Table (3.3.2) the power of the test should not be adversely affected.

4.3.3 Two sample t-test

Let $\underline{X} = (X_1,\ldots,X_n)$ and $\underline{Y} = (Y_1,\ldots,Y_n)$ be independent random samples of size n from normal populations X and Y with means μ_X, μ_Y and variances σ^2_X, σ^2_Y respectively. Let (X_{Ri},Y_{Ri}) be the rounded values of (X_i,Y_i) corresponding to a rounding lattice with interval of width w and lattice position c.

For testing the hypothesis $H_0:\mu_X = \mu_Y$ vs $H_1:\mu_X \neq \mu_Y$ assuming $\sigma^2_X = \sigma^2_Y$ the test statistic is given by (3.3-13). Under H_1

$$t = \frac{(\bar{X}-\bar{Y}) - d}{S_P\sqrt{\frac{2}{n}}} \sim t_{2n-2}(\delta) \qquad (4.3\text{-}7)$$

where $t_{2n-2}(\delta)$ is a non-central t distribution with non-centrality parameter $\delta = \frac{d}{\sigma}\sqrt{\frac{n}{2}}$, with $d = \mu_X - \mu_Y$, the difference between μ_X and μ_Y under H_1. Under rounding

$$t_R = \frac{(X_R-Y_R) - d}{\sqrt{\frac{S^2_{XR} + S^2_{YR}}{n}}} \qquad (4.3\text{-}8)$$

where $\bar{X}_R$, $\bar{Y}_R$, S^2_{XR} and S^2_{YR} are defined as in (3.3-13).

Under H_0, for values of r within the recommended ranges given in Table (3.3.6), the distribution of t_R (3.3-13) closely approximates that of a t distribution. Under H_1 we would expect a similar situation, namely that the distribution of t_R (4.3-8) is in close agreement with a non-central t distribution for the recommended values

of r. Using the same approach as in the single t–test, the non–centrality parameter under rounding will be

$$\delta_R \simeq \frac{\sigma}{\sqrt{\left[1 + \frac{r^2}{12}\right]}} < \delta \qquad (4.3\text{-}9)$$

where δ is the non–centrality parameter of the corresponding non–central t distribution for unrounded data. As for the one sample t–test, rounding will result in the test (4.3–7) being less powerful. Using an approach similar to that with the one sample t–test, it can be shown that if r satisfies the recommendation given in Table (3.3.6), then a reasonable approximation to the distribution of t_R (4.3–8) is given by $t(\delta_R)$.

Table 4.3.8

Range of values of P_R at α = 0.05, mean and variance of t_R when n = 10 and r = 2.0 for a two sample t–test

P	P_R (simulation)	$P_R(t(\delta_R))$	$E[t_R]$		$V[t_R]$	
	lower tail	lower tail	simulation	$t(\delta_R)$	simulation	$t(\delta_R)$
0.30	0.239-0.253	0.248	1.04-1.08	1.06	1.09-1.18	1.13
0.50	0.397-0.418	0.411	1.52-1.57	1.55	1.10-1.30	1.19
0.70	0.577-0.597	0.589	2.00-2.09	2.04	1.12-1.50	1.28
0.95	0.879-0.889	0.883	3.01-3.20	3.09	1.21-1.41	1.30

Table 4.3.9

Range of values of P_R at $\alpha = 0.05$, mean and variance of t_R when n = 25 and r = 2.0 for a two sample t-test

P	P_R (simulation lower tail	$P_R(t(\delta_R))$ lower tail	$E[t_R]$ simulation	$t(\delta_R)$	$V[t_R]$ simulation	$t(\delta_R)$
0.30	0.246-0.250	0.253	1.00-1.01	1.00	1.03-1.06	1.05
0.50	0.409-0.414	0.411	1.50-1.51	1.47	1.04-1.08	1.06
0.70	0.588-0.594	0.595	1.92-1.95	1.94	1.04-1.11	1.07
0.95	0.883-0.888	0.883	2.90-3.00	2.94	1.06-1.20	1.12

Tables (4.3.8) and (4.3.9) show a selection of results for lower tail tests where $\alpha = 0.05$. Results for the upper tail were very similar.

There is clearly close agreement between the simulated results and those obtained from the distribution of $t(\delta_R)$. The values of the mean and variance indicate that the distribution of $t(\delta_R)$ will closely approximate that of t_R. This is also evident from the close agreement between the P_R values from simulation and the distribution of $t(\delta_R)$. To obtain the P_R values from the distribution of $t(\delta_R)$, tables given by Owen (1965) were used.

Some values of P_R were also obtained for $\alpha = 0.01$ and 0.001. Close agreement between the distributions of $t(\delta_R)$ and t_R was also found for these two levels of significance, if the value of r was in the recommended range. It is reasonable to conclude that if the value of r is in the recommended range given in Table (3.3.6)

the distribution of $t(\delta_R)$ will be a good approximation to t_R.

Simulation Results

In section (3.3.3) the recommended range of r for which the significance level of the test was found acceptable under rounding for $\alpha = 0.05$ (one tailed) was $r \leq 2.0$ for n = 5, 10 and 25. For $\alpha = 0.05$ values of P_R obtained by simulation for n = 5, 10 and 25, where r = 2.0 and 1.5, are given in Table (4.3.10) for lower tail tests. P_R values for upper tail tests were very similar.

Table 4.3.10

Range of values of P_R for a two-sample t-test for $\alpha = 0.05$ (lower tail)

	n = 5		n = 10		n = 25	
P	r = 2.0	r = 1.5	r = 2.0	r = 1.5	r = 2.0	r = 1.5
0.30	0.261-0.267	0.266-0.272	0.239-0.253	0.265-0.270	0.246-0.250	0.267-0.269
0.50	0.426-0.435	0.444-0.447	0.397-0.418	0.441-0.449	0.409-0.414	0.444-0.448
0.70	0.605-0.613	0.632-0.636	0.577-0.597	0.631-0.638	0.588-0.594	0.635-0.637
0.95	0.890-0.894	0.914-0.916	0.879-0.889	0.916-0.918	0.883-0.888	0.911-0.914
α = 0.05	0.051-0.054	0.051-0.053	0.049-0.053	0.050-0.050	0.049-0.050	0.050-0.051

The P_R values are consistent across the range of n. The only exception is n = 5 at r = 2.0, where the power levels are slightly greater than expected. This has been caused by the more discontinuous nature of t_R, which is a result of the small sample size and rounding as coarse as r = 2.0.

For the ranges of r given in Table (3.3.6) the distribution of t_R can be approximated by $t(\delta_R)$. Tables or a suitable approximation can be used to estimate the power P_R [refer to section 4.3.1].

4.3.4 F test for equality of two variances

Let $\underline{X} = (X_1,\ldots,X_n)$ and $\underline{Y} = (Y_1,\ldots,Y_n)$ be independent random samples of size n from normal populations X and Y with means μ_X, μ_Y and variances σ_1^2, σ_2^2 respectively. Let (X_{Ri},Y_{Ri}) be the rounded values of (X_i,Y_i) corresponding to a rounding lattice with interval of width w and lattice position c.

For testing the hypothesis $H_0:\sigma_1^2 = \sigma_2^2$ vs $H_1:\sigma_1^2 \neq \sigma_2^2$, the test statistic is given by (3.3-17). Under H_1

$$F = \frac{S_X^2\,\sigma_2^2}{S_Y^2\,\sigma_1^2} \sim F_{n-1,n-1} \qquad (4.3\text{-}10)$$

where $F_{n-1,n-1}$ is an F distribution with (n-1,n-1) degrees of freedom.

Without any loss of generality we assume that σ_2^2 is fixed and $\sigma_1^2 = \theta\sigma_2^2$. For testing the hypothesis $H_0:\theta = 1$ vs $H_1:\theta \neq 1$, under rounding the test statistic is

$$F_R = \frac{S_{XR}^2}{S_{YR}^2} \qquad (4.3\text{-}11)$$

where S_{XR}^2 and S_{YR}^2 are defined as in (3.3-16).

For any given sample of rounded data, w and c are fixed. Under H_1 the degrees of precision of the X_{Ri} and Y_{Ri} values are $r_1 = w/\sigma_2\sqrt{\theta}$ and $r_2 = w/\sigma_2$ respectively. The normal populations X and Y have been rounded to differing degrees of precision. This is as in section (3.3.4) where the F statistic was considered for differing precision in the two populations. We can again make use of the work of Gayen (1950) and Sheppard's corrections of the mean and variance of the test statistic F_R (4.3-11). These are given by

$$E[F_R] \simeq \left[\frac{12+r_1^2}{12+r_2^2}\right] E\left[F\right]_{H_1}, \quad V[F_R] \simeq \left[\frac{12+r_1^2}{12+r_2^2}\right]^2 V\left[F\right]_{H_1} \qquad (4.3\text{-}12)$$

where $E[F]_{H_1}$ and $V[F]_{H_1}$ are the mean and variance of F (4.3-10) and $r_1 = r_2/\sqrt{\theta}$.

The first two moments of F_R indicate that if $r_1 > r_2$, then rounding will cause the test statistic to have an increased mean and variance. For $r_1 < r_2$ there will be a decrease in the mean and variance. The change in the mean will be the most important factor indicating the effect of rounding on the power of the test. The degree to which rounding will change the mean of the distribution of F is controlled by the factor

$$B = \left[\frac{12+r_1^2}{12+r_2^2}\right] \qquad (4.3\text{-}13)$$

The value of B will depend on r_1 and r_2. In

(i) lower tailed tests, as $\theta < 1$, $r_1 > r_2$ and $B > 1$. Rounding will cause the distribution of F to shift to the right, resulting in a reduction in power.

(ii) upper tailed tests, as $\theta > 1$, $r_1 < r_2$ and $B < 1$. Rounding will cause the distribution of F to shift to the left, resulting in a reduction in power.

For a fixed level of power the shift in the distribution of F will be greatest in the lower tailed tests. Hence we would expect rounding to cause a greater reduction in the power for this type of test.

Simulation Results

In section (3.3.4) the recommended ranges of r for which the significance level of the test was found acceptable under rounding for $\alpha = 0.05$ (one tailed) were $r \leq 1.0$ when $n = 5$, $r \leq 1.5$ when $n = 10$ and 25. For $\alpha = 0.05$, values of P_R obtained by simulation for $n = 5$, 10 and 25 are shown in Tables (4.3.11 to 4.3.13).

Table 4.3.11

Range of values of P_R for a F-test for $\alpha = 0.05$, $n = 5$, $r = 0.5$ and 1.0

P	r = 0.5		r = 1.0	
	lower tail	upper tail	lower tail	upper tail
0.30	0.285-0.287	0.294-0.296	0.197-0.268	0.285-0.287
0.50	0.464-0.467	0.493-0.495	0.241-0.473	0.477-0.478
0.70	0.641-0.643	0.695-0.696	0.253-0.691	0.678-0.681
0.95	0.806-0.924	0.948-0.949	0.254-0.943	0.943-0.944

Table 4.3.12

Range of values of P_R for a F–test for $\alpha = 0.05$, n = 10, r = 0.5, 1.0 and 1.5

P	r = 0.5		r = 1.0		r = 1.5	
	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
0.30	0.288-0.290	0.293-0.294	0.258-0.260	0.276-0.279	0.120-0.285	0.254-0.258
0.50	0.474-0.477	0.490-0.492	0.409-0.414	0.467-0.468	0.142-0.487	0.431-0.436
0.70	0.664-0.666	0.690-0.689	0.545-0.587	0.665-0.667	0.150-0.700	0.630-0.634
0.95	0.921-0.923	0.946-0.947	0.669-0.889	0.937-0.938	0.152-0.940	0.924-0.925

Table 4.3.13

Range of values of P_R for a F–test for $\alpha = 0.05$, n = 25, r = 0.5, 1.0 and 1.5

P	r = 0.5		r = 1.0		r = 1.5	
	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
0.30	0.291-0.293	0.292-0.295	0.263-0.267	0.275-0.278	0.208-0.250	0.241-0.253
0.50	0.481-0.482	0.490-0.492	0.430-0.434	0.461-0.464	0.307-0.414	0.414-0.424
0.70	0.675-0.676	0.688-0.687	0.604-0.610	0.658-0.659	0.401-0.591	0.608-0.618
0.95	0.933-0.935	0.946-0.947	0.874-0.88	0.932-0.934	0.515-0.904	0.905-0.973

As expected the P_R values show that the effect of rounding is greatest in the lower tail. For the maximum value of r allowed in the recommended range of r, the power can be considerably reduced. At these maximum values of r, the lattice effect c can result in the P_R values having a wide range. For fixed r, increasing

the size of n was seen to reduce the effect of rounding on the power. For example comparing the P_R values for r = 1.0, n = 5, with those for r = 1.0, n = 10 clearly shows this.

The values of P_R have indicated that the power of the F–test can be adversely affected by rounding if the recommended ranges of r for α = 0.05 given in Table (3.3.8) are applied. The results from the simulation suggest that a better recommendation for α = 0.05, which would give more acceptable levels of power, is:

$n = 5 \quad r \leq 0.5$

$n = 10 \quad r \leq 0.5$

$n = 25 \quad r \leq 1.0$

In this section values of P_R have been generated only for one tailed tests where α = 0.05. However, the results clearly indicate that the recommended ranges of r in Table (3.3.8) are unsuitable with respect to the level of power. In the light of the power results for α = 0.05, 'safer' recommended values of r that may be regarded as acceptable for n = 5, 10 and 25 are given in Table (4.3.14).

Table 4.3.14

Values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25 in a F-test for equality of two variances

α(%)	One tailed						Two tailed
	0.1/1.0/5.0		1.0/5.0		5.0		5.0
n	LT	UT	LT	UT	LT	UT	
5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5
10	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5	r ≤ 0.5
25	r ≤ 1.0	r ≤ 1.0	r ≤ 1.0	r ≤ 1.0	r ≤ 1.0	r ≤ 1.0	r ≤ 1.0

4.3.5 Analysis of Variance (ANOVA)

In the one-way ANOVA, for testing the hypothesis $H_0:\alpha_i = 0$ vs $H_1:\alpha_i \neq 0$ (i=1,...,k), the test statistic is given by (3.3-25). Under H_1

$$F = \frac{Q_1/(k-1)}{Q_2/(nk-k)} \sim F_{k-1,k(n-1)}(\varphi) \qquad (4.3\text{-}15)$$

where $F_{k-1,k(n-1)}(\varphi)$ is a non-central F distribution with k-1 and k(n-1) degrees of freedom and non-centrality parameter $\varphi = \sqrt{\frac{n\Sigma\alpha_i^2}{k\sigma^2}}$ where $\alpha_i = \mu-\mu_i$ and μ and μ_i respectively the overall mean and ith population mean.

Let all the k samples be drawn from normal populations which have been rounded according to the same rounding lattice, with rounding interval w and lattice position c. The test statistic under H_1 for rounded data is

$$F_R = \frac{Q_{1R}/(k-1)}{Q_{2R}/(nk-k)} \qquad (4.3\text{-}16)$$

where Q_{1R} and Q_{2R} are defined as in (3.3–26).

Under H_0 for values of r within the recommended ranges given in Table (3.3.10) the distribution of F_R (3.3–26) closely approximated that of an F distribution. Under H_1 we would expect a similar situation. Namely that the distribution of F_R (4.3–16) will be in close agreement with a non–central F distribution for the recommended values of r. As the normal populations are subject to rounding the non–centrality parameter will be

$$\varphi_R = \sqrt{n \sum_i \frac{\alpha_{iR}}{k\sigma_R^2}}$$

where α_{iR} is the ith sample effect under rounding and σ^2_R the variance of the rounded normal populations. The effect α_{iR} is simply the difference $\mu_{iR}-\mu_R$, where μ_R and μ_{iR} are respectively the overall mean and ith population mean for rounded data. By applying Sheppard's corrections to approximate μ_R, μ_{iR} and σ^2_R, an estimate of φ_R is given by:

$$\varphi_R = \sqrt{n \sum_i \frac{\alpha_{iR}}{k\sigma_R^2}} = \sqrt{n \sum_i \frac{(\mu_R-\mu_{iR})^2}{k\sigma_R^2}}$$

$$\simeq \sqrt{\frac{n \sum_i (\mu-\mu_i)^2}{k\sigma^2\left[1 + \frac{r^2}{12}\right]}}$$

$$\simeq \frac{\varphi}{\sqrt{1 + \frac{r^2}{12}}} < \varphi \qquad (4.3\text{-}17)$$

Hence from (4.3-17), we would expect under H_1 the distribution of F_R to be in close agreement with a non-central F distribution, $F_{k-1,k(n-1)}(\varphi_R)$, where φ_R is less than φ. This reduction in the non-centrality parameter caused by rounding will result in the one-way ANOVA becoming less powerful for rounded data. Using an approach similar to that for the t-tests, it can be shown that if the value of r satisfies the recommendation given in Table (3.3.10), then the distribution of F_R is approximate to $F_{k-1,k(n-1)}(\varphi_R)$. Table (4.4.15) and (4.4.16) show a selection of results. They illustrate the close agreement between the simulated results and those obtained from the distribution of $F_{k-1,k(n-1)}(\varphi_R)$. The values of the mean and variance indicate that the distribution of F_R will closely approximate that of $F_{k-1,k(n-1)}(\varphi_R)$. This is also evident from the close agreement of the P_R values for $\alpha = 0.05$, from the simulation and the non-central F distribution. To obtain the P_R values from the distribution of $F_{k-1,k(n-1)}(\varphi_R)$ tables given by Tiku (1967) were used.

Table 4.3.15

Range of values of P_R at $\alpha = 0.05$, mean and variance of F_R for k = 5, n = 5 and r = 2.0 in a one-way analysis of variance

P	P_R (simulation)	$P_R[F(\varphi_R)]$	$E[F_R]$		$V[F_R]$	
			simulation	$F(\varphi_R)$	simulation	$F(\varphi_R)$
0.30	0.218-0.240	0.248	2.04-2.09	2.07	2.35-2.53	2.44
0.50	0.367-0.384	0.378	2.77-2.80	2.77	3.67-3.98	3.71
0.70	0.544-0.561	0.548	3.60-3.68	3.61	4.94-6.38	5.50
0.95	0.856-0.863	0.855	5.71-5.88	5.90	9.11-12.01	11.10

Table 4.3.16

Range of values of P_R at $\alpha = 0.05$, mean and variance of F_R for $k = 3$, $n = 25$ and $r = 2.0$ in a one-way analysis of variance

P	P_R (simulation)	$P_R[F(\varphi_R)]$	$E[F_R]$		$V[F_R]$	
			simulation	$F(\varphi_R)$	simulation	$F(\varphi_R)$
0.30	0.231-0.238	0.233	2.13-2.16	2.14	3.50-3.67	3.58
0.50	0.383-0.390	0.385	2.98-3.04	3.00	5.36-5.72	5.54
0.70	0.560-0.570	0.569	4.07-4.17	4.13	7.82-8.49	8.16
0.95	0.865-0.871	0.869	7.08-7.29	7.18	14.83-15.64	15.64

Some values of P_R were also obtained for $\alpha = 0.01$ and 0.001. Close agreement between the distributions of $F_{k-1,k(n-1)}(\varphi_R)$ and F_R was also found for these two levels of significance, if the value of r was in the recommended range. It is reasonable to conclude that if the value of r is in the recommended range given in Table (3.3.10) the distribution of $F_{k-1,k(n-1)}$ will be a good approximation to F_R.

Simulation Results

In section (3.3.5) the recommended ranges of r for which the significance level of the test was found acceptable under rounding for $\alpha = 0.05$ were $r \leq 1.5$ when $N = 15$ and $r \leq 2.0$ when $N \geq 16$. For $\alpha = 0.05$ values of P_R were obtained by simulation for a selection of N and r values. Table (4.3.17) contains some of these results.

Table 4.3.17

Range of values of P_R at $\alpha = 0.05$ for $k = 3$, $n = 5$, 25 and $k = 5$, $n = 5$ in a one-way analysis of variance

k×n	3×5	3×25		5×5	
P	r = 1.5	r = 2.0	r = 1.5	r = 2.0	r = 1.5
0.30	0.247-0.250	0.231-0.238	0.256-0.259	0.218-0.240	0.249-0.253
0.50	0.421-0.423	0.383-0.390	0.426-0.431	0.367-0.384	0.423-0.426
0.70	0.609-0.611	0.560-0.570	0.617-0.620	0.544-0.561	0.611-0.614
0.95.	0.904-0.905	0.865-0.871	0.904-0.906	0.856-0.863	0.900-0.903

Although Table (4.3.17) shows only a selection of the P_R values obtained, they faithfully represent the entire body of results.

For the ranges of r given in Table (3.3.10) the distribution of F_R can be approximated by $F_{k-1,k(n-1)}(\varphi_R)$. Hence tables or a suitable approximation to the non-central F distribution can be used to estimate the power P_R. For example tables given by Tiku (1967) or an approximation to the cumulative distribution of the non-central F may be obtained by a method outlined by Norton (1983).

From section (3.3.5) the significance levels in a one-way or two-way ANOVA were found to be very similar for rounded data. We would expect the same situation in the case of the power. For the simulations (Appendix B) the P_R values were in close agreement for the one and two-way ANOVA.

4.3.6 Compensation for Rounding

The chi-squared test statistic has been found to be the least robust to rounding. Both the significance level and power of the test were found to be very sensitive to rounding. This sensitivity was a result of the mean and variance of the test statistic being increased for rounded data. The increases in the mean and variance are approximately of the same order for all n; thus it may be possible to use a standard correction to the test statistic to compensate for the rounding effect. If the denominator in (3.3-8) is changed to $\sigma_0^2 + w^2/12$ then the adjusted test statistic, denoted by χ^2_C is

$$\chi^2_C = \frac{(n-1)S_R^2}{\sigma_0^2 + \frac{w^2}{12}} \qquad (4.3\text{-}18)$$

From (3.3-10) the approximate mean and variance of the distribution of χ^2_C are respectively (n-1) and 2(n-1).

This adjustment to the test statistic will compensate for the effect of rounding. For large enough n (4.3-18) will be in closer agreement with a chi-squared distribution than is (3.3-8). For small n, the distribution of χ^2_C may have too many discontinuities and the adjusted test statistic may not be very effective in compensating for the rounding process. To study the effectiveness of the adjusted test statistic, values of α_R and P_R were obtained by simulation for various combinations of (n,r).

Simulation Results

For $n \leq 5$ the distribution of χ^2_C was found to be in poor agreement with a chi–squared sistribution. The probability of $\chi^2_C = 0$ was still high for this size of n and resulted in the α values being severely distorted by rounding. For $n > 5$ the χ^2_C test statistic gave a closer agreement between the α and α_R values than did χ^2_R (3.3–8). Table (4.3.18) shows a selection of results. In this table the values of α_R for χ^2_C and χ^2_R test statistics are compared.

Table 4.3.18

Minimum and maximum values of $\alpha_R(\%)$ found for 11 values of c in the χ^2_R and χ^2_C test statistics

n = 10

r	test statistic		α(%) lower tail 0.1	1.0	2.5	5.0	α(%) upper tail 5.0	2.5	1.0	0.1
1.5	χ^2_C	min	0.05	0.58	0.58	2.72	3.06	2.19	0.69	0.03
		max	0.24	2.11	2.11	5.43	3.76	2.54	0.95	0.07
	χ^2_R	min	0.05	0.58	0.58	0.58	11.58	4.76	2.52	0.40
		max	0.24	2.11	2.11	2.11	12.84	6.04	2.77	0.43
1.0	χ^2_C	min	0.06	0.99	3.32	5.04	5.03	2.21	1.04	0.09
		max	0.10	1.07	3.33	5.26	5.18	2.35	1.06	0.09
	χ^2_R	min	0.06	0.44	2.05	3.32	6.72	3.74	1.75	0.23
		max	0.09	0.60	2.47	3.33	6.75	3.77	1.77	0.23

n = 25

r	test statistic		α(%) lower tail 0.1	1.0	2.5	5.0	α(%) upper tail 5.0	2.5	1.0	0.1
1.5	χ^2_C	min	0.02	0.33	2.13	3.68	4.80	2.27	0.85	0.09
		max	0.20	0.95	2.90	4.81	5.21	2.47	0.94	0.10
	χ^2_R	min	0.00	0.06	0.33	1.31	15.24	9.71	4.81	0.85
		max	0.06	0.37	0.95	2.17	15.46	10.72	5.21	0.94
1.0	χ^2_C	min	0.09	0.09	2.22	4.83	4.94	2.36	0.90	0.09
		max	0.13	0.98	2.55	5.21	5.12	2.39	1.04	0.11
	χ^2_R	min	0.05	0.57	1.43	2.88	8.74	5.04	2.23	0.30
		max	0.07	0.63	1.72	3.17	9.39	5.34	2.35	0.32

Table (4.3.18) shows that the difference between α and α_R values is less for the χ^2_C test statistic. This is especially true in the upper tail. As n increased in value, χ^2_C became a better approximation to a chi-squared distribution with n-1 degrees of freedom. This is apparent by comparing the values of α and a_R for n = 10 and 25 in Table (4.3.18).

In using the adjusted test statistic χ^2_C, it was possible to extend the range of r for which the significance level of the test will be acceptable. Table (4.3.19) shows the values of r for which α_R was found acceptable for n = 5, 10 and 25.

Table 4.3.19

The values of the degree of precision r that may be regarded as acceptable for n = 5, 10 and 25 in a chi-squared test for a variance using the χ^2_C test statistic

	One tailed test						Two tailed test
α(%)	0.1/1.0/5.0		1.0/5.0		5.0		5.0
n	LT	UT	LT	UT	LT	UT	
5	$r \leq 0.5$	$r \leq 0.25$	$r \leq 0.5$	$r \leq 0.25$	$r \leq 0.5$	$r \leq 0.5$	$r \leq 0.5$
10	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$
25	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$	$r \leq 1.0$

In comparing Table (3.3.4) and (4.3.19), using χ^2_C allows the range of r to be extended to 1.0 for n = 10 and 25. For χ^2_C, values of P_R are shown in Table (4.3.20) for α = 0.05 (one tailed), for r = 1.0 where n = 10 and 25. These values of P_R indicate what loss in power to expect if the maximum recommended value of r is used for n = 10 and 25. For lower values of r, the power

reduction will be less.

Table 4.3.20

Range of values of P_R at α = 0.05 for the test statistics χ^2_C for r = 1.0 where n = 10 and 25

	n = 10 r = 1.0		n = 25 r = 1.0	
P	lower tail	upper tail	lower tail	upper tail
0.30	0.237-0.274	0.276-0.278	0.259-0.274	0.270-0.275
0.50	0.389-0.446	0.464-0.465	0.424-0.444	0.457-0.459
0.70	0.554-0.614	0.662-0.663	0.604-0.619	0.654-0.657
0.95	0.836-0.882	0.936-0.937	0.880-0.883	0.932-0.933

4.4 Discussion and Conclusions

In Chapter 3 recommended ranges or r were given for n = 5, 10 and 25, in which the significance level of the test may be considered acceptable. It is natural to investigate what the power of the test will be using these recommendations. Although the power of each test was considered mainly for α = 0.05, it provided a clear indication of the level of power we should expect under rounding.

For hypothesis tests concerned with means, rounding resulted in a loss of power. The main result of this section was that by adjusting the non-centrality parameter in the non-central distribution, an estimate of the power under rounding could be obtained. For values of r for which the significance level was found acceptable, this estimate of power for rounded data was found to be reasonably accurate. This estimate of power can be useful in practice, in providing an idea of the expected loss in power under rounding.

For the one sample t–test, for values of r for which the significance level was found acceptable, the power is still of an appreciable magnitude. However for the two sample t–test, one and two–way analysis of variance, a lower value of r than that recommended for the significance level may be necessary to limit the loss in power caused by rounding. Although for the analysis of variance only two layouts were considered, the results suggest the likely level of power to expect with such statistical procedures for rounded data.

Unlike hypothesis tests for means, those concerned with variances will not have a constant degree of precision r under H_1. This was found to make the power more sensitive to rounding. With the chi–squared test its lack of robustness to rounding meant that acceptable levels of significance were obtainable only for low values of r. For these low values of r, the power was not found to be adversely affected for $\alpha = 0.05$. In general, the results indicate that the chi–squared test will be only slightly less powerful for values of r for which the significance level was found acceptable. However for the F–test, the reduction in the power could be severe for values of r for which the significance level was found acceptable. In order to maintain a more suitable level of power the recommended ranges for r had to be reduced.

The chi–squared test could be made more robust to rounding by making a simple adjustment to the test statistic. Test of hypothesis regarding the value of σ^2 should be based on the adjusted test statistic, χ^2_C. By using this adjusted test statistic it was possible to extend the range of r for which the level of significance of the test is acceptable. However by extending the range of r there will be a corresponding loss in power.

In this chapter we have confined our attention to values of $n \leq 25$. However the power levels for $n = 25$ provide a good indication of what level of power to expect for larger sample sizes.

The results of Chapter 4 have provided a 'good feel' for the robustness of the tests considered in Chapter 3, with respect to power. More importantly we now know what level of power a test may have for the values of r for which the significance level was found acceptable.

CHAPTER 5

THE EFFECT OF ROUNDING ON THE SIGNIFICANCE LEVEL AND POWER OF CERTAIN NORMAL TEST STATISTICS FOR NON-NORMAL DATA

5.1 Introduction

In Chapters 3 and 4 the effect of rounding on the significance level and power of certain test statistics was considered, for an underlying normal population. However in many situations the statistical tests in Chapter 3 must be used when the assumption of normality is invalid. There has been much research on the robustness of these tests when the population is non-normal. To date there has been no study of the possible effects of rounding on a statistical test when the assumption of normality is invalid. The following illustrates how non-normality may increase the effect of rounding on a statistical test.

Chapter 2 showed that the normal distribution is very robust to rounding with respect to its moments. For example the maximum error in the population moments μ, σ^2, $\sqrt{\beta_1}$ and β_2 are less than $10^{-4}\sigma$, 0.2σ, 10^{-2} and 10^{-1} respectively for rounding as coarse as $r = 1.5$, when the population is normal. However, for non-normal populations the situation can change, as shown by the contour diagrams in Chapter 2. Increased skewness and kurtosis in a population can result in a greater rounding error in the population moments. The moments of the sampling distribution of a test statistic depend on the moments of the parent population. Any change in the population moments caused by rounding will directly affect the sampling distribution of the test statistic. For example how will this greater effect on the moments of non-normal populations caused by rounding be reflected in the significance level of a test? This chapter aims to indicate how much 'non-normality' can be allowed without the effect of rounding seriously distorting the significance level and power of the tests in Chapter 3.

Chapter 3 considered only test statistics where the population is assumed to be normal. However in section (5.4) a paper by Tricker (1984a) is reviewed, which investigates the effect of rounding on a test statistic where the population is assumed to be exponential.

5.2 Description of the Investigation

As in Chapter 2, the family of Johnson distributions is taken to represent the family of non–normal distributions, and we use the same set of 29 Johnson distributions as were used by Pearson and Please (1975). The four added to this set in Chapter 2 to represent U shaped distributions will not be considered. This study deals with only moderate departures from normality; where the population is very non–normal one may argue that such tests as those in Chapter 3 should not be used.

In this present study the significance and power level were evaluated for each statistical test in Chapter 3. The following results were obtained by simulation, with sample sizes of n = 10 and 25:

(i) for each Johnson distribution, the significance level of the test under rounding was evaluated for values corresponding to the lower and upper 5% points under normal theory conditions, with no rounding;

(ii) for a selection of Johnson distributions, the power of the test statistic under rounding was evaluated for values of the alternative hypothesis H_1 corresponding to powers 0.3 and 0.7 under normal theory conditions with $\alpha = 0.05$.

The usual lattice positions c = −0.5, −0.4, ..., 0.4, 0.5 were considered for each value of r. These eleven lattice values will indicate the effect of the position of the rounding lattice on the significance level and power of a test.

The value of α = 0.05 was chosen as this is normally the first level of significance at which the null hypothesis is rejected. To keep the computing within reasonable bounds, the study of the power was restricted. The power was evaluated at a 'low' and 'high' level for α = 0.05 for a selection of Johnson distributions. Sample sizes larger than 25 were not considered, as there is then a reasonable chance that non−normality can be detected and some corrective action taken.

The emphasis of this chapter is to provide guidance on what happens to the significance level and power of the tests in Chapter 3, if the values of r which were recommended for normal populations are applied when the population is non−normal. Essentially how far can the degree of precision r recommended for normal populations be applied to the non−normal situation? Hence the values of r considered will be in the vicinity of those which were recommended for sample sizes 10 and 25, when the population is normal. In this study the following approaches were used:

(a) The effect of rounding on the significance level of a test

(i) Approximations to the sampling moments of the test statistics were examined to provide a rough outline of what characteristics are to be expected when sampling from rounded non−normal data.

(ii) Estimation of the sampling distribution of the test statistic for rounded data was obtained by Monte Carlo methods.

(b) The effect of rounding on the power of a test

As in Chapter 4 for normal populations, only Monte Carlo methods were used to estimate the power of a test for rounded non-normal data.

In this investigation the significance level and power for each test for samples drawn from unrounded Johnson distributions were also obtained. Pearson and Please (1975) in their work on the robustness of normal test statistics for unrounded data, considered the same set of Johnson distributions used in this study. They used simulation to consider the effect of non-normality on the significance level of a test statistic. The results were presented in the form of charts.

In addition to test statistics investigated by Pearson and Please, the one and two-way analysis of variance are considered. Also the robustness of a test statistic is looked at from both a level of significance and power aspect for unrounded data.

To perform the necessary analysis two previous FORTRAN programs were adapted. The programs SIMUL and PSIMUL were modified to allow samples to be drawn from Johnson distributions as well as from normal distributions. As the significance and power level for each test for samples drawn from unrounded Johnson distributions were also required, a further program USIMUL was written.

The results of the SIMUL, USIMUL and PSIMUL programs were based on 10,000 iterations, which gave adequate precision for the 0.05 level of significance, and the 0.3, 0.7 levels of power. Of course the results obtained by simulation will be

subject to sampling errors. However, these errors will be small for 10,000 iterations.

Quality of Results

The SIMUL and PSIMUL programs were tested to check validity of their results. As both programs were adapted from earlier programs only the generation of deviates from the Johnson distributions had to be checked. The USIMUL program was checked by comparing the results with those of Pearson and Please (1975). Although their results were in diagrammatic form, there was no apparent disagreement between these results and those given by the USIMUL program. A final check was established by comparing the results from SIMUL and PSIMUL programs, where the Johnson distributions are not subject to rounding, with the results from the USIMUL program.

5.3 Test Statistics

In this section it is assumed without any loss of generality that in the null case (H_0) the Johnson distributions have a mean of zero and variance one. The non-null case (H_1) was handled by adjusting the parameters in the standardised distributions to give the required power under normal theory conditions. Throughout this section α will denote the level of significance of the test for samples drawn from normal populations subject to no rounding, while α_J and α_{JR} will be the resulting levels of significance of the test where the samples have been drawn fron non-rounded and rounded Johnson populations respectively. Thus α_J and α_{JR} are simply the probabilities that the test statistic fell above or below the α significance level limits for non-rounded and rounded data respectively. A

similar notation will be used for the power of the test. P will denote the level of power for samples drawn from normal populations subject to no rounding, while P_J and P_{JR} will be the resulting power of the test where the samples have been drawn from non-rounded and rounded Johnson populations respectively.

The results are presented as follows for each test statistic:

(i) Before discussing the results, approximate expressions are given for the moments of the test statistic for a Johnson population that has been subject to rounding. These approximate moments help to indicate any possible changes in the distribution of the test statistic under rounding.

(ii) The level of significance α_{JR} will be discussed for $\alpha = 0.05$, where the sample sizes are n = 10 and 25.

A diagram will show the values of (β_1,β_2) that may be regarded as acceptable for a given value of r at n = 10 and 25, where $\alpha = 0.05$. When using a single tailed test, values of (β_1,β_2) will be considered acceptable if the significance level for unrounded normal data is 5%, while for rounded Johnson data with degree of precision r, α_{JR} lies between 3%-7%. The values of r considered will be those that were recommended when the underlying population is normal (Chapter 3). Such a diagram allows the reader to know the range of β_1 and β_2 values that may be used for the values of r recommended for normal populations, while keeping the level of significance within reasonable bounds. Essentially the diagrams provide guidance on how far the degree of precision r recommended for normal populations can be

applied to the non-normal situation, with respect to the level of significance.

(iii) The power values P_{JR} will be discussed for a selection of Johnson distributions for n = 10 and 25, where α = 0.05. The main purpose of this section will be to investigate if the level of power for values of (β_1,β_2) which gave a level of significance α_{JR} between 3%-7% is adversely affected by rounding.

Appendix C contains tables showing the values of

(i) α_{JR} for the entire (β_1,β_2) plane for n = 10 and 25 where α = 0.05;

(ii) P_{JR} for a selection of Johnson distributions for n = 10 and 25, where α = 0.05.

These tables will allow the reader to determine the general trend of what happens to the significance and power level for a test for non-normal data which has been subject to rounding. For a list of all output produced by the SIMUL, PSIMUL and USIMUL programs in the study, refer to Appendix C.

5.3.1 One sample t-test

Let $\underline{X} = (X_1,\ldots,X_n)$ be a random sample of size n from a Johnson population X, with shape parameters $\sqrt{\beta_1}$ and β_2. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the rounded value of X_i corresponding to a rounding lattice with an interval of width w and lattice position c. For testing the hypothesis H_0: $\mu = \mu_0$ the t-test statistic is given by

$$t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \qquad (5.3\text{-}1)$$

where $X \sim J(\sqrt{\beta_1},\beta_2)$ ie a Johnson distribution with shape parameters $\sqrt{\beta_1}$ and β_2.

As we have assumed that the Johnson distributions have mean zero and variance one, then under rounding (5.3-1) becomes

$$t_R = \frac{\bar{X}_R}{S_R/\sqrt{n}} \qquad (5.3\text{-}2)$$

where $\bar{X} = \sum_i \frac{X_i}{n}$, $S^2 = \frac{1}{n-1}\sum_i (X_i-\bar{X})^2$, $\bar{X}_R = \sum_i \frac{X_{Ri}}{n}$, $S_R^2 = \frac{1}{n-1}\sum_i (X_{Ri}-\bar{X}_R)^2$

To be able to use Geary (1947) results to obtain an approximation to the moments of t_R as in section (3.3.1), the population means of X_R and X must be equal. As illustrated in Chapter 2 for non-normal populations the means of X_R and X are often unequal. However, there are methods which can be used to find the approximate mean and variance of some function of random variables. [Details of this method are given in Appendix C]. Approximations to the first two moments of t_R (5.3-2) for large n are given by:

$$E[t_R] \approx \frac{\mu_R\sqrt{n}}{\sigma_R}\left[1 + \frac{3}{8}(\beta_{2R}-1)/n\right] - \frac{1}{2\sqrt{n}}\sqrt{\beta_{1R}}$$

$$V[t_R] \approx 1 + \frac{1}{4}\frac{\mu_R^2}{\sigma_R^2}(\beta_{2R}-1) - \frac{\mu_R}{\sigma_R}\sqrt{\beta_{1R}} \qquad (5.3\text{-}3)$$

where μ_R, σ^2_R, $\sqrt{\beta_{1R}}$ and β_{2R} are the rounded parameters of X_R. [Proof in Appendix C].

If the Johnson populations are not subject to rounding we have from Geary (1947)

$$E[t] = -\frac{1}{2\sqrt{n}}\sqrt{\beta_1} - O(n^{-3/2})$$

$$V[t] = 1 + \frac{1}{4}(8+7\beta_1)/n + O(n^{-2}) \qquad (5.3\text{-}4)$$

where μ, σ^2, $\sqrt{\beta_1}$ and β_2 are the parameters of X.

Although (5.3–3) and (5.3–4) indicate that the variance of the test statistic will change under rounding, the change in the mean is more important, as it will cause the greatest effect on the significance levels. As we have set $\mu = 0$ under H_0, μ_R represents the change in the population mean due to rounding. It is this value of μ_R that is crucial in determining how the mean alters. The contour diagram Figure (2.2.10) shows that the change will be greatest in the top right hand corner of the (β_1,β_2) region, where the departure from symmetry is greatest. Hence we would expect the significance level α_J to be more affected by rounding in the top right hand corner of the (β_1,β_2) region. This effect will not diminish as n increases.

Simulation Results : Significance Levels

In Appendix C, Tables (C.1) and (C.2) show the range in values of α_J and α_{JR} when n = 10 and 25, for $\alpha = 0.05$. The results presented in these two tables provide a good evaluation of what happens to the significance level of a one sample t–test for rounded non–normal data. The results speak for themselves and the reader may determine the general trend of what happens to the level of significance for a given r over the (β_1,β_2) region.

For Johnson distributions not subject to rounding, the simulation results are in agreement with the charts of Pearson and Please (1975). In Table (C.1) when

n = 10, the results show how increasing population skewness draws the lower and upper tail α_J values in opposite directions. For n = 25 (Table C.2) the shift in the significance levels is much less than for n = 10. For neither value of n does the population kurtosis β_2 produce much effect.

As suggested by (5.3-3) the greatest change in the α_J values caused by rounding is in the top right hand corner of the (β_1,β_2) region. This coincides with the region where the rounding error in the population mean is largest and departure from symmetry in the distribution X greatest. Increasing n to 25 generally results in rounding having less effect on the α_J values, except in the top right hand corner of the (β_1,β_2) region, where the effect is slightly greater. This suggests that where the departure from symmetry is severe, increasing the sample size may not diminish the effect of rounding; in fact it may increase it. Generally for increasing skewness the lattice effect (c) caused the range in the α_{JR} values to widen. This was especially noticeable for $\beta_2 \leq 2.4$ and $\sqrt{\beta_1} \geq 0.6$.

In a normal population, with α = 0.05 (one tailed) the recommended ranges of r for n = 10 and 25 were respectively $r \leq 1.0$ and $r \leq 1.5$. For values of r in these ranges, Figures (5.3.1) and (5.3.2) show the values of (β_1,β_2) where the α_J and α_{JR} values lie between 3%-7%. The figures are so presented to allow a comparison to be made between that region of (β_1,β_2) plane for which the significance level lies between 3%-7% for unrounded data and that region of (β_1,β_2) plane for which the significance level lies between 3%-7% for rounded data. This provides guidance on how far the degree of precision r recommended for a normal population can be applied to non-normal situations, without making the test less robust with respect to the level of significance.

Figure 5.3.1: Region in (β_1,β_2) plane where $\alpha_J[J]$ and $\alpha_{JR}(X)$ values lie between 3%–7% in a one sample t–test for n = 10 and α = 0.05 (one tailed), where r = 1.0 and 0.5.

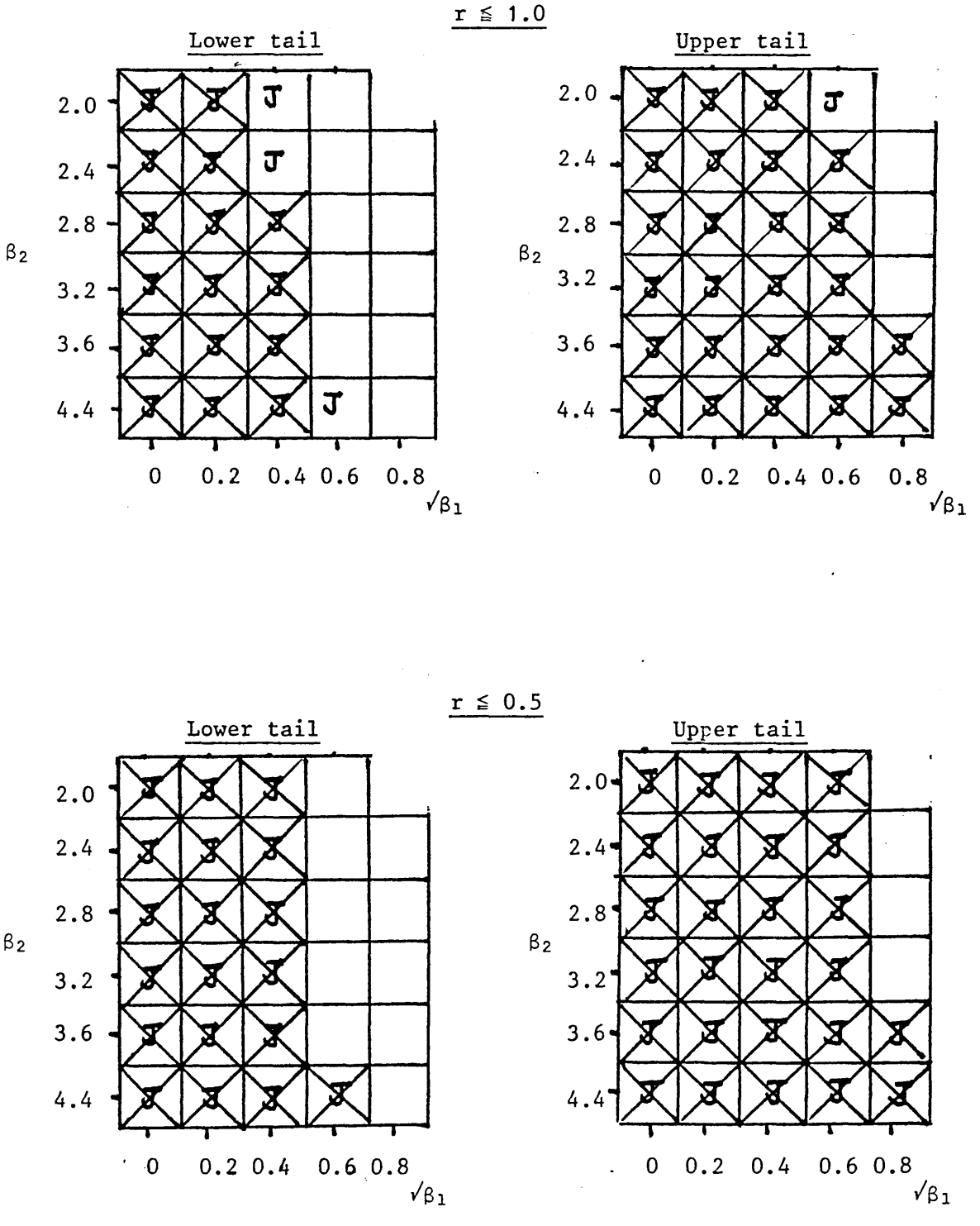
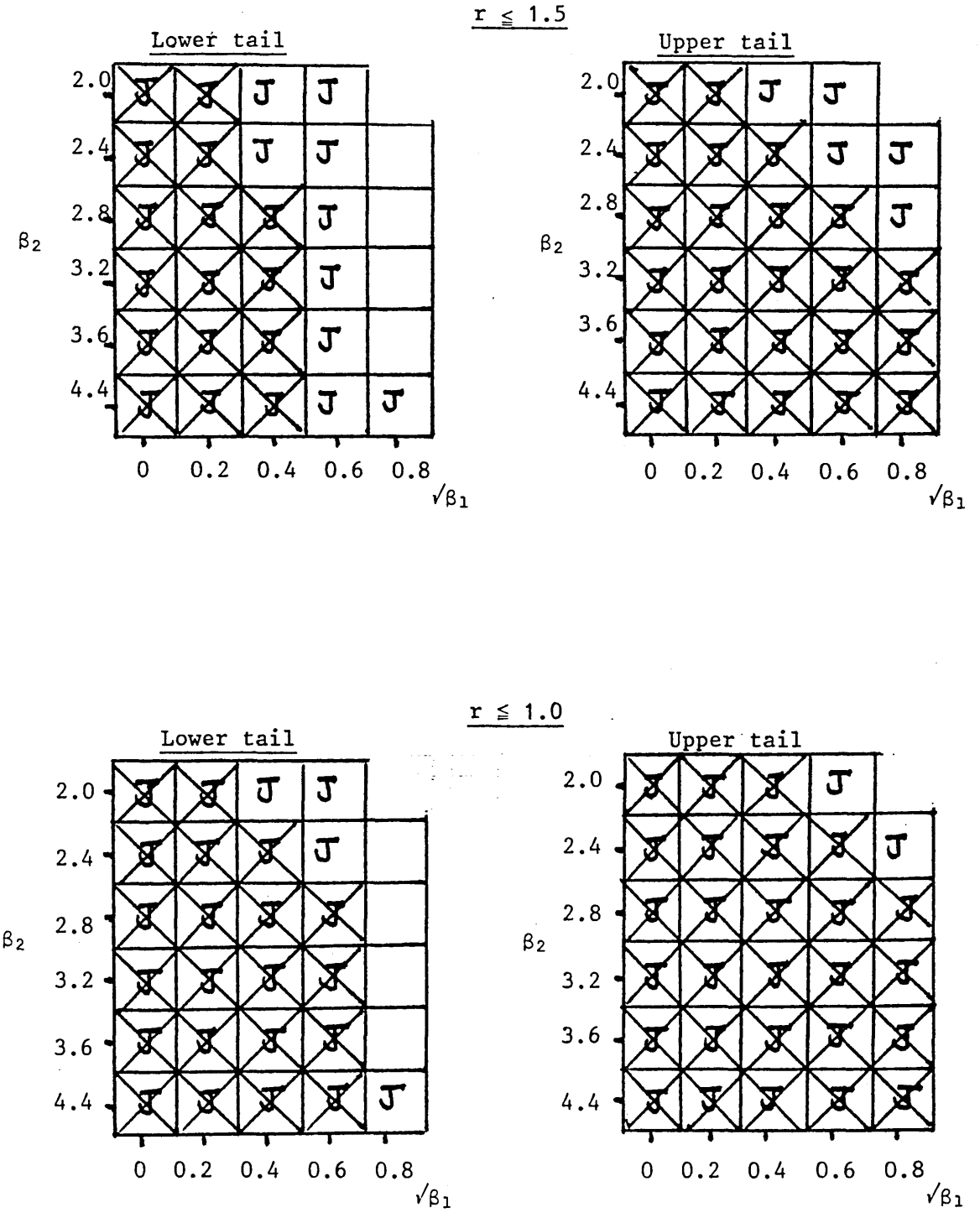


Figure 5.3.2: Region in (β_1,β_2) plane where α_J [J] and α_{JR} (X) values lie between 3%–7% in a one sample t–test for n = 25 and α = 0.05 (one tailed), where r = 1.5 and 1.0.



The striking feature from Figures (5.3.1) and (5.3.2) is the similarity of the regions in the (β_1,β_2) plane where α_J and α_{JR} lies between 3%–7%. For both n = 10 and 25, where $r \leq 1.0$ these regions are almost identical.

Simulation Results : Power Levels

In Appendix C Table (C.3) shows the range of values of P_J and P_{JR} when n = 10, for P = 0.3 and 0.7, where β_2 = 2.0, 2.4 and 4.4.

The power of the one sample t–test has been studied by such authors as Ghurye (1949), Srivastava (1958) and Posten (1978), the main conclusions being that positive skewness causes a reduction in power in the region of low power and increase in power in the region of high power, for upper tail tests. For lower tail tests there is a reverse in the situation. The results in Table (C.3) for Johnson distributions not subject to rounding exhibit this pattern.

The changes in the P_J values due to rounding were similar to those found for the significance levels α_J. The greatest change occurred in the top right hand corner of the (β_1,β_2) region. For n = 25 the power levels P_J and P_{JR} were of similar magnitude as for n = 10 except in the region $\beta_2 \leq 2.4$ and $\sqrt{\beta_1} \geq 0.6$. In this region for $r \geq 1.0$ the distortion in the P_J values caused by rounding was noticeably greater for n = 25. In general the results indicated that for n = 10, $r \leq 1.0$ and n = 25, $r \leq 1.5$ for values of (β_1,β_2) which gave levels of significance α_{JR} between 3%–7%, values of P_J are not adversely affected by rounding. Table (C.3) shows a selection of power levels P_{JR} for n = 10.

5.3.2 Chi-squared test for a variance

Let $\underline{X} = (X_1,\ldots,X_n)$ be a random sample of size n from a Johnson population X, with shape parameters γ and β_2. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the rounded value of X_i corresponding to a rounding lattice with interval of width w and lattice position c. For testing the hypothesis $H_0:\sigma^2 = \sigma_0^2$ the chi-squared test statistic is given by

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad (5.3\text{-}5)$$

where $X \sim J(\gamma,\beta_2)$.

As we have assumed the Johnson distributions have mean zero and variance one, under rounding (5.3-5) becomes:

$$\chi_R^2 = (n-1)S_R^2 \qquad (5.3\text{-}6)$$

where S^2 and S^2_R are defined as in (5.3-2).

From (3.3-10) the exact first and second moments of χ^2_R are given by

$$\begin{aligned} E[\chi_R^2] &= \sigma_R^2(n-1) \\ V[\chi_R^2] &= \sigma_R^4\left[2(n-1)+\left[1-\frac{1}{n}\right]^2(\beta_{2R}-3)\right] \end{aligned} \qquad (5.3\text{-}7)$$

where σ^2_R and β_{2R} are the variance and kurtosis of a rounded Johnson distribution. If the Johnson population is not subject to rounding we have

$$E[\chi^2] = (n-1)$$

$$V[\chi^2] = \left[2(n-1) + \left[1 - \frac{1}{n}\right]^2(\beta_2-3)\right] \qquad (5.3\text{-}8)$$

Study of the moments suggests that the values of σ^2_R and β_{2R} will control how the mean and variance of the test statistic will alter for rounded data. For a normal population these values could be approximated by Sheppard's corrections (3.3–4). As shown in section (2.2.3) this is not necessarily so for non–normal populations. The contour diagrams Figures (2.2.11) and (2.2.13) show that the change in σ^2 and β_2 can vary considerably over the (β_1,β_2) plane for rounded data. However except for the top right hand corner of the (β_1,β_2) plane there will be reasonable agreement between the values of σ^2_R and β_{2R} and those obtained for a normal population. We would then expect the effect of rounding on the significance levels to be similar for normal and non–normal populations, this effect being that rounding causes the lower tail values to decrease while the upper tail values will increase. In the top right hand corner of (β_1,β_2) plane where departure from symmetry is greatest, the parameters σ^2_R and β_{2R} can be considerably different in value than for a normal population. In this region of (β_1,β_2) we would expect the effect of rounding on the significance levels to be different than under normality.

Simulation Results : Significance Levels

Tables (C.4) and (C.5) show the range in values of α_J and α_{JR} for n = 10 and 25, for α = 0.05.

For Johnson distributions not subject to rounding, the simulation results are in agreement with the charts of Pearson and Please (1975). In Tables (C.4) and (C.5) the results for α_J show that for $\beta_2 < 3$ there is less than expected in the tails and for $\beta_2 > 3$ there is more than expected. The change in the significance levels is much the same for all values of $\sqrt{\beta_1}$ and always present for n = 10 and 25. As n increases, so does the disagreement between the values of α and α_J.

As expected for a large area of the (β_1,β_2) plane rounding causes the α_J values in the lower and upper tails to respectively decrease and increase. [Tables C.4 and C.5]. Rounding had the same effect as when the population is normal (section 3.3.2). Only when $\beta_2 \leq 2.4$ and $\sqrt{\beta_1} \geq 0.6$ is this not necessarily so. In this region of the (β_1,β_2) plane the α_{JR} values are more erratic in behaviour and the lattice effect (c) considerably greater. Generally rounding resulted in a greater inbalance between the lower and upper tails, this being worse as r and n increased in size.

In a normal population with $\alpha = 0.05$ (one tailed) the recommended range of r for n = 10 and 25 was $r \leq 0.5$. For values of r in this range Figure (5.3.3) shows the values of (β_1,β_2) where for $\alpha = 0.05$, the α_J and α_{JR} values lie between 3% to 7%.

Figure 5.3.3 Region in (β_1,β_2) plane where $\alpha_J[J]$ and $\alpha_{JR}(X)$ values lie between 3%–7% in a chi-squared test for a variance for n = 10 and 25, where α = 0.05 and r < 0.5

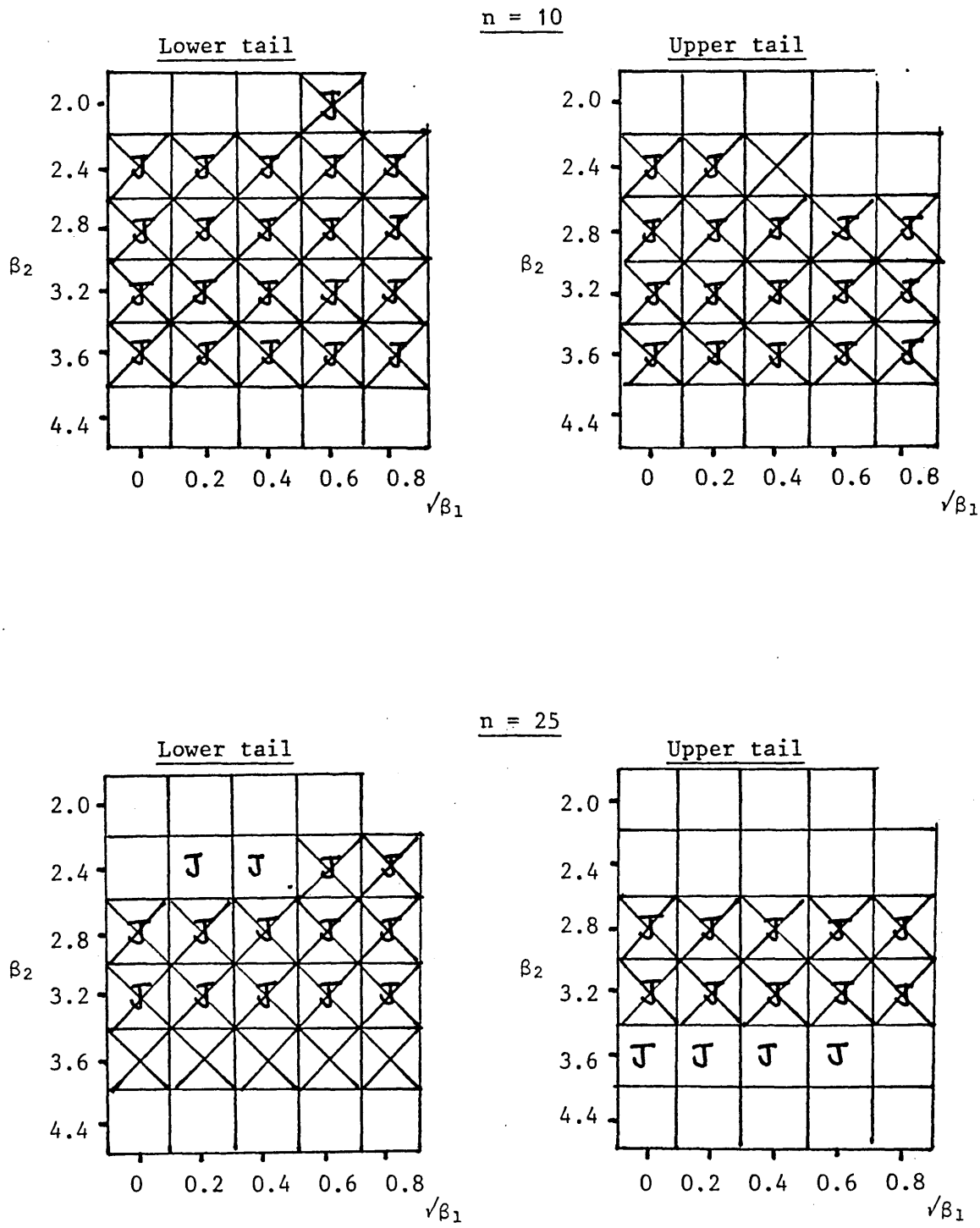


Figure (5.3.3) shows that for n = 25 the similarity of the regions in the (β_1,β_2) plane where α_J and α_{JR} lie between 3%–7% is less than for n = 10. The reason is that as n increases in size rounding causes the agreement between α_J and α_{JR} to deteriorate.

Simulation Results : Power Levels

Simulation results for the power were obtained for β_2 = 2.0, 2.4, 3.6 and 4.4, when n = 10 and 25. Table (C.6) shows the range in values of P_J and P_{JR} when n = 10, for P = 0.3 and 0.7, where β_2 = 2.4 and 3.6.

When the Johnson distributions are not subject to rounding the power values P_J showed a tendency to be less than expected in the lower tail and more than expected in the upper tail, for $\beta_2 < 3$. For $\beta_2 > 3$ the simulation was in reverse.

For both H_0 and H_1 the distribution of the test statistics are essentially the same. Thus we would expect rounding to have the same effect on the test statistic under H_0 and H_1, this effect being to shift the distribution to the right. As expected this caused the P_J values in the lower and upper tails to respectively decrease and increase. As with the significance levels this may not be so in the top right hand corner of the (β_1,β_2) plane. The simulated values of P_{JR} for n = 10 and 25 showed this to be so, as is evident from the selection of results in Table (C.6) for n = 10.

In general the results show that if the degree of precision r recommended for a normal population for n = 10 and 25 ($r \leq 0.5$) is used, there will not be a large

difference between the significance and power levels for unrounded and rounded data. The only exception to this may be for $\beta_2 \leq 2.4$ and $\sqrt{\beta_1} \geq 0.6$.

5.3.3 Two sample t-test

Let $\underline{X} = (X_1,\ldots,X_n)$ and $\underline{Y} = (Y_1,\ldots,Y_n)$ be independent random samples of size n from Johnson populations X and Y, with shape parameters $\sqrt{\beta_1}$ and β_2. The means and variances of X and Y are μ_X, μ_Y and σ^2_X, σ^2_Y respectively. Let (X_{Ri},Y_{Ri}) be rounded values of (X_i,Y_i) corresponding to a rounding lattice with interval of width w and lattice position c.

For testing the hypothesis $H_0:\mu_X = \mu_Y$, the t-test statistic is given by

$$t = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_X^2+S_Y^2}{n}}} \qquad (5.3\text{-}9)$$

where X and Y $\sim J(\sqrt{\beta_1},\beta_2)$

and under rounding

$$t_R = \frac{\bar{X}_R-\bar{Y}_R}{\sqrt{\frac{S_{XR}^2+S_{YR}^2}{n}}} \qquad (5.3\text{-}10)$$

where $\bar{X} = \sum_i \frac{X_i}{n}$, $\bar{Y} = \sum_i \frac{Y_i}{n}$, $S_X^2 = \frac{1}{n-1}\sum_i (X_i-\bar{X})^2$, $S_Y^2 = \frac{1}{n-1}\sum_i (Y_i-\bar{Y})^2$,

$\bar{X}_R = \sum_i \frac{X_{iR}}{n}$, $\bar{Y}_R = \sum_i \frac{Y_{iR}}{n}$, $S_{XR}^2 = \frac{1}{n-1}\sum_i (X_{iR}-\bar{X}_R)^2$, $S_{YR}^2 = \sum_i (Y_{iR}-\bar{Y}_R)^2$

To obtain approximations to the first four moments of t_R (5.3-10) we use expansions for the moments of a two sample t-test statistic, for samples drawn

from non-normal populations from Geary (1947).

$$
\begin{aligned}
E[t_R] &= 0 \\
V[t_R] &= 1 + \frac{1}{n} + O(n^{-2}) \\
\sqrt{\beta_1}(t_R) &= 0 \\
\beta_2(t_R) &= 3 + \frac{3}{n} + O(n^{-2})
\end{aligned}
\qquad (5.3\text{-}11)
$$

To order n^{-1} the expressions (5.3–11) are also the first four moments of the test statistic t (5.3–9), where the Johnson distributions are not subject to rounding. It follows that we expect there to be little difference between the distribution of t and t_R for non-normal populations. In fact the expressions (5.3–11) are also the first four moments of the test statistic for normal rounded data (3.3–15). This indicates that there will be very little difference between the distributions of the test statistic for rounded normal or non-normal data.

Simulation Results : Significance Levels

There is no difference between the upper and lower tail significance levels, as the distribution of the test statistic is symmetrical about zero. The α_J and α_{JR} values for n = 10 and 25 were very similar. Table (C.7) shows the range in values of α_J and α_{JR} when n = 10 for α = 0.05, only for the lower tail. Table (C.8) shows the range in the level of significance (α_R) for rounded normal data.

For Johnson distributions not subject to rounding the simulated values of α_J for n = 10 and 25, were in agreemeent with the charts of Pearson and Please (1975). For the entire region of the (β_1,β_2) plane there was close agreement between the

α_J and α values. This is illustrated by the values of α_J for n = 10 given in Table (C.7). As concluded by Pearson and Please (1975), where the two sample sizes are equal, the two sample t-test is very robust to departures from normality, with respect to level of significance.

As the approximate moment calculations had led us to expect, there was a very close agreement between the α_J and α_{JR} values for very coarse rounding. As expected an increase in n from 10 to 25 caused a slight improvement in this agreement. The results given in Table (C.7) for n = 10 illustrate the close agreement found between the significance levels for rounded and unrounded data from Johnson populations.

The striking feature of the two sample t-test is that departures from normality make very little difference to the effect that rounding has on the significance level. The significance levels for rounded normal data (α_R) and rounded non-normal data (α_{JR}) were very similar. Comparison of the results in Tables (C.7) and (C.8) illustrate this point.

Where the populations are normal, for α = 0.05 (one tailed) the recommended range of r for n = 10 and 25 was $r \leq 2.0$. For values of r in this range Figure (5.3.4) shows the values of (β_1,β_2) where the α_J and α_{JR} values lie between 3%-7%. For both n = 10 and 25, where $r \leq 2.0$, the regions in the (β_1,β_2) plane for rounded and unrounded data are identical. This would also be true if the interval 3%-7% was reduced to 4%-6%.

Simulation Results : Power Levels

For the two sample t-test simulation results for the power were obtained for the entire (β_1,β_2) plane for n = 10 and 25. Table (C.9) shows the range in values of P_J and P_{JR} for n = 10, for P = 0.3 and 0.7, where β_2 = 2.0, 2.5 and 4.4, for lower tail. Values of P_J and P_{JR} for the upper tail were very similar.

Posten (1978) carried out an extensive computer simulation study of the effect of non-normality on the power levels of a two sample t-test, where the sample sizes are equal, his conclusions being that, departures from normality have very little effect on the power levels of the test. The P_J value found confirmed this, as illustrated by a selection of results given in Table (C.9).

Before discussing the power levels for rounded data, an explanation is required on their behaviour under normality.

Under H_0 the population means μ_X and μ_Y are assumed equal. This means that for a given r and c, the positions of the distributions X and Y on the rounding lattice are identical. Thus under H_0 the effect of rounding on the parameters of the populations X and Y will be the same. However for H_1 this is not necessarily so. Under H_1 the population means μ_X and μ_Y are unequal and the positions of the distributions X and Y on the rounding lattice may be different. For a given r and c this may cause rounding to have a different effect on the parameters of the populations X and Y. This means that rounding can cause the difference in the population means under H_1 to change thus resulting in a change in the alternative hypothesis. When the departure from normality is moderate, the position of the distribution on the rounding lattice will have little effect on the population means.

When the departure is more extreme, the effect can be considerably greater. The more extreme distributions occur in the top right hand corner of the (β_1,β_2) plane. It is in this region where the change in the value of the alternative hypothesis caused by rounding will become significant. This change in the alternative hypothesis will result in less agreement between the P_J and P_{JR} values.

Although the significance levels changed very little for rounded data this was not so for the power of the test. Generally in the (β_1,β_2) plane the effect of rounding on the power was greater than when the populations were normal. The results showed that increasing skewness increased the range of the P_{JR} values, this being especially so for $\beta_2 < 3$. As expected the distortion in the P_J values caused by rounding was very noticeable in the top right hand corner of the (β_1,β_2) plane. [Table C.9]. For n = 25 the power levels P_J and P_{JR} were of a similar magnitude as for n = 10, except in the region $\beta_2 \leq 2.4$ and $\sqrt{\beta_1} \geq 0.6$. In this region for $r \leq 1.5$ the distortion in the P_J values caused by rounding was greater for n = 25. Except for $(\sqrt{\beta_1},\beta_2)$ = (0.6,2.0) and (0.8,2.4), for $r \leq 1.0$ the power for rounded normal and non-normal populations was similar. As an illustration Figure (5.3.5) indicates the region in the (β_1,β_2) plane where the level of power is in excess of 0.25 and 0.60 respectively for P = 0.3 and 0.7, for n = 10 and 25, where $r \leq 1.5$ and $r \leq 1.0$. Figure (5.3.5) shows that for rounding as coarse as r = 1.5, for over 75% of the (β_1,β_2) plane the power is still in excess of 0.25 and 0.60 respectively for P = 0.3 and 0.7.

Figure 5.3.4 Region in (β_1,β_2) plane where $\alpha_J[J]$ and $\alpha_{JR}(X)$ values lie between 3%–7% in a two sample t–test for n = 10 and 25,

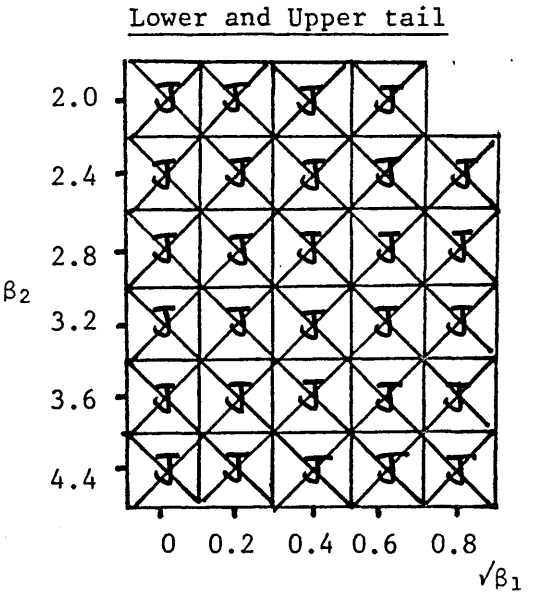
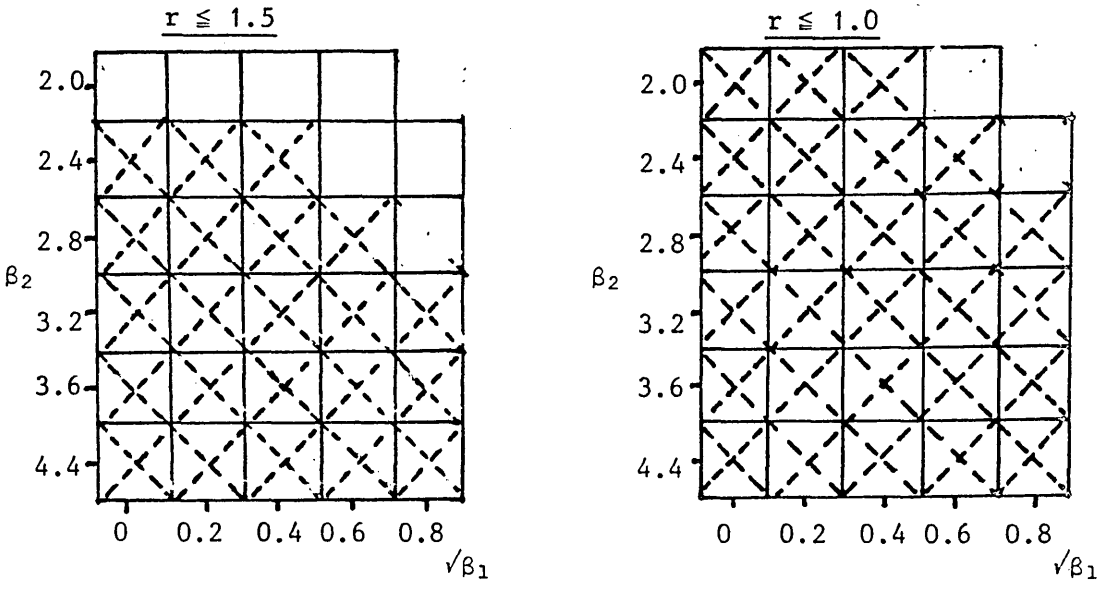


Figure 5.3.5 Region in (β_1,β_2) plane where $P_{JR}(\cdot)$ is in excess of 0.25 and 0.60 respectively for P = 0.3 and P = 0.7, for a two sample t–test where n = 10 and 25, when α = 0.05



5.3.4 F–test for equality of two variances

Let $\underline{X} = (X_1,\ldots,X_n)$ and $\underline{Y} = (Y_1,\ldots,Y_n)$ be independent random samples of size n from Johnson populations X and Y, with shape parameters $\sqrt{\beta_1}$ and β_2. The means and variances of X and Y are μ_X, μ_Y and σ^2_X, σ^2_Y respectively. Let (X_{Ri},Y_{Ri}) be the rounded values of (X_i,Y_i) corresponding to a rounding lattice with interval of width w and lattice position c.

For testing the hypothesis $H_0:\sigma^2_X = \sigma^2_Y = \sigma^2$, the F–test statistic is given by

$$F = S^2_X/S^2_Y \qquad (5.3\text{-}12)$$

and under rounding

$$F = S^2_{XR}/S^2_{YR} \qquad (5.3\text{-}13)$$

where S^2_X, S^2_Y, S^2_{XR} and S^2_{YR} are defined as in (5.3-10).

From (3.3–19), approximations to the first two moments of F_R are

$$\begin{aligned} E[F_R] &= 1 + \frac{1}{n}(\beta_{2R}-1) + O(n^{-2}) \\ V[F_R] &= \frac{2}{n}(\beta_{2R}-1) + O(n^{-2}) \end{aligned} \qquad (5.3\text{-}14)$$

where β_{2R} is the measure of kurtosis for the populations.

If the Johnson populations are not subject to rounding we have

$$E[F] = 1 + \frac{1}{n}(\beta_2 - 1) + O(n^{-2})$$

$$V[F] = \frac{2}{n}(\beta_2 - 1) + O(n^{-2}) \qquad (5.3\text{-}15)$$

Study of the moments suggests that the difference between the values of β_2 and β_{2R} will be important in determining how the mean and variance of the test statistic will alter for rounded data. The contour diagram Figure (2.2.13) shows how this difference can vary considerably over the (β_1, β_2) plane under rounding. It will be the size of this difference which is important in determining how the α_J values are distorted by rounding.

Simulation Results: Significance Levels

Tables (C.10) and (C.11) show the range in values of α_J and α_{JR} when n = 10 and 25 for α = 0.05.

For Johnson distributions not subject to rounding the simulation results are in agreement with the charts of Pearson and Please (1975). In Tables (C.10) and (C.11) the results for α_J show that for $\beta_2 < 3$ there is less than expected in the tails and for $\beta_2 > 3$ there is more than expected. This change in the significance levels is much the same for all values of $\sqrt{\beta_1}$ and always present for n = 10 and 25. As n increases so does the agreement between the α and α_J values.

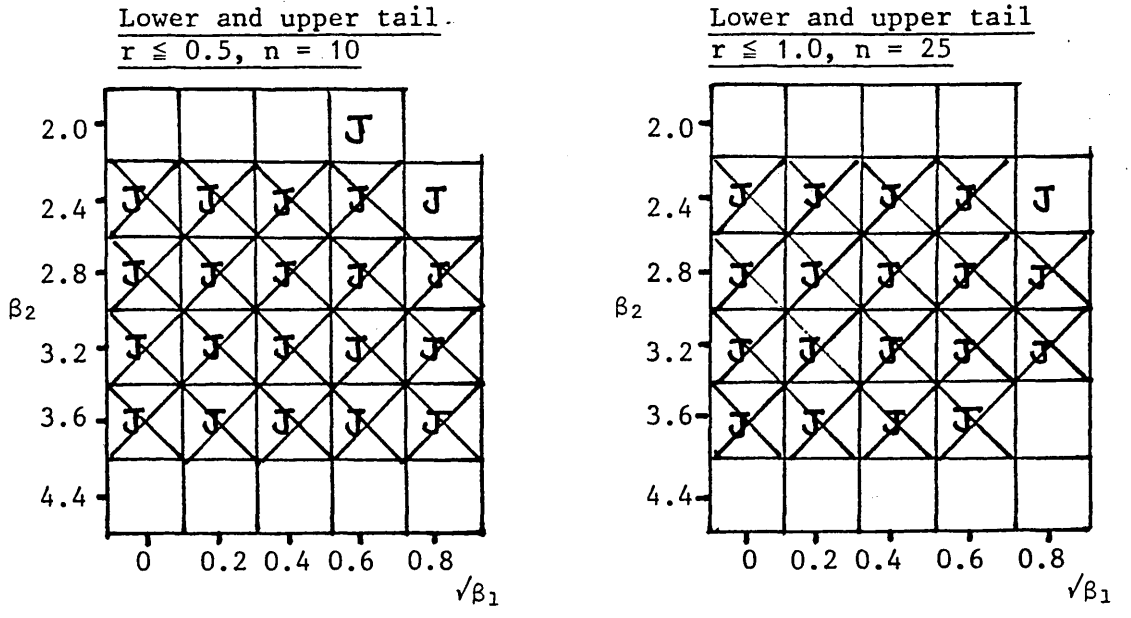
As suggested by the approximate moments, the size of the difference between β_2 and β_{2R} was important in determining how the α_J values were influenced by rounding. The contour diagram Figure (2.2.13) shows the maximum difference

between β_2 and β_{2R}, over the (β_1,β_2) place. Comparing this contour diagram with the results in Tables (C.9) and (C.10) shows a strong relationship between this maximum difference and the amount of distortion in the α_J values due to rounding.

For $\beta_2 < 3$ the distortion in the α_J values due to rounding was generally greater than for $\beta_2 > 3$. Rounding has the most effect in the top right hand corner of the (β_1,β_2) plane. For fixed r, increasing population skewness increases the effect of rounding, as shown by the wider range in the α_{JR} values. [Tables C10 and C11]. For $r \leq 1.0$ when n = 10 and 25 there was reasonable agreement between the α_J and α_{JR} values except in the region $\beta_2 < 3$ and $\sqrt{\beta_1} \geq 0.6$.

In a normal population, the recommended ranges of r given in Table (3.3.8) for the level of significance were adjusted because of the level of power (section 4.3.4). The modified values of r that may be regarded as acceptable with respect to level of significance are given in Table (4.3.14). For $\alpha = 0.05$ (one tailed) the recommended values of r for n = 10 and 25 from this table are $r \leq 0.5$ and $r \leq 1.0$ respectively. For values of r in these ranges Figure (5.3.6) shows the values of (β_1,β_2) where for $\alpha = 0.05$, the α_J and α_{JR} values lie between 3% to 7%. For both n = 10 and 25 the regions in the (β_1,β_2) plane where α_J and α_{JR} lie between 3%–7% are almost identical.

Figure 5.3.6: Region in (β_1,β_2) plane where $\alpha_J[J]$ and $\alpha_{JR}(X)$ values lie between 3%–7% in a F–test for equality of variances for n = 10 and 25, where α = 0.05.



Simulation Results : Power Levels

For the F–test, simulation results for the level of power were obtained for β_2 = 2.4, 2.8, 3.2 and 3.6 where n = 10 and 25. Tables (C.12) and (C.13) show the range in values of P_J and P_{JR} when n = 10 and 25 respectively, for P = 0.3 and 0.7. Table (C.13) gives power levels for only the lower tail.

For Johnson distributions not subject to rounding the results indicated that for $\beta_2 < 3$, there was a reduction in power in the region of low power (P = 0.3) and an increase in power in the region of high power (P = 0.7). For $\beta_2 > 3$ the situation is the reverse. The P_J values are much the same for all values of $\sqrt{\beta_1}$.

As with normal populations (section 3.3.4), rounding will cause the distribution of the test statistic under H_1 to shift to the right or left. This will result in a reduction in power. As indicated by the selection of results in Tables (C.12) and (C.13) rounding results in a similar effect on power levels as that found for normality, this effect being to reduce power in both tails, with the lower tail having the greater reduction. The greatest change in P_J values due to rounding is in the top right hand corner of the (β_1,β_2) plane.

When n = 10 and r $\leq$ 0.5, for values of (β_1,β_2) which gave levels of significance α_{JR} between 3%–7% the P_J values were not adversely affected by rounding. This is illustrated by a selection of results given in Table (C.12). When n = 25 and r $\leq$ 1.0, for values of (β_1,β_2) which gave levels of significance between 3%–7% the P_J values were only distorted severely by rounding for lower tail tests when β_2 = 2.4 and $\sqrt{\beta_1} \geq 0.4$ [Table C.13].

5.3.5 Analysis of Variance (ANOVA)

One-way Analysis of Variance - Fixed Effects Model

The structure we shall assume for the one-way layout fixed effect model is that given by (3.3-25). Let the samples be drawn from Johnson distributions with shape parameters $\sqrt{\beta_1}$ and β_2. We shall assume that all the Johnson distributions have been rounded to the same rounding lattice, with rounding interval w and lattice position c. For testing the hypothesis $H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_R$, the test statistic for rounded data is given by

$$F_R = \frac{Q_{1R}/(k-1)}{Q_{2R}/(nk-k)} \qquad (5.3\text{-}16)$$

where $Q_{1R} = \sum_{i=1}^{n} n(\bar{X}_{Ri.} - \bar{X}_{R..})^2$, $Q_{2R} = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{Rij} - \bar{X}_{Ri.})^2$

Again using the results of Gayen (1950) we have

$$E[F_R] = 1 + \frac{2}{N} + O(N^{-2})$$

$$V[F_R] = \frac{2}{k-1} + \frac{2}{N(k-1)} [5+k-(\beta_{2R}-3)] + O(N^{-2}) \qquad (5.3\text{-}17)$$

where $N = nk$ and β_{2R} is the measure of kurtosis for the rounded Johnson populations.

If the Johnson populations are not subject to rounding we have

$$E[F] = 1 + \frac{2}{N} + O(N^{-2})$$

$$V[F] = \frac{2}{k-1} + \frac{2}{N(k-1)}\,[5+k-(\beta_{2R}-3)] + O(N^{-2}) \qquad (5.3\text{-}18)$$

Study of the moments suggests that rounding will cause the variance to change, the change depending on the difference between β_2 and β_{2R}. Although for non-normal distributions this difference can vary considerably over the (β_1,β_2) plane, its effect will diminsh as N increases.

It follows that we would expect the distributions of F and F_R to be very similar for non-normal populations.

Simulation Results : Significance Levels

Significance levels α_J and α_{JR} were obtained for k = 3 and 5, when n = 10 and 25, for α = 0.05. A selection of these results is in Table (C.14) for k = 3 and n = 10.

The effects of non-normality on the level of significance of the F-test in the one way ANOVA has been studied by several authors. For example, investigations by Pearson (1931), Geary (1947) and Grayen (1950) have indicated that the significance level of the F-test is very insensitive to non-normality of the parent populations. The simulated values α_J confirm this. The values of α_J given in Table (C.14) illustrate how robust the F-test is to departures from non-normality with respect to level of significance.

As the approximate moment calculations have led us to expect, there was reasonable agreement between the α_J and α_{JR} values for rounding as coarse as r = 2.0. As k or n increased there was a slight improvement in the agreement between the α_J and α_{JR} values. The results in Table (C.14) illustrate the type of agreement found between the significance levels for rounded and unrounded data for the ANOVA layouts considered. In the one way ANOVA, departures from normality made very little difference to the effect rounding had on the significance level. The significance levels for rounded normal data (α_R) and rounded non-normal data (α_{JR}) were found to be similar. Comparison of the results in Tables (C.14) and (C.15) illustrates this point. When the populations are normal, for α = 0.05 the recommended range of r for nk $\geq$ 16 was r $\leq$ 2.0. For the one way layouts considered the α_J and α_{JR} values were found to lie between 3%-7% for the entire (β_1,β_2) plane. This would also be true if the interval 3%-7% was reduced to 4%-6%.

Simulation Results : Power Levels

Power levels P_J and P_{JR} were obtained for k = 3 and 5, when n = 10 and 25, for α = 0.05. Table (C.18) shows the range in values of P_J and P_{JR} for k = 3 and n = 10, where P = 0.3 and 0.7.

The effect of violation of the normality assumption on the power in the analysis of variance has been studied by several authors. Kanji (1976, 1977) used simulation to study the effects of non-normality on the power in analysis of variance for both one and two way layouts. David and Johnson (1951) considered the extent to which non-normality affects the F-test. Their findings indicate that the inferences concerning means the power calculated under normal theory is only slightly affected

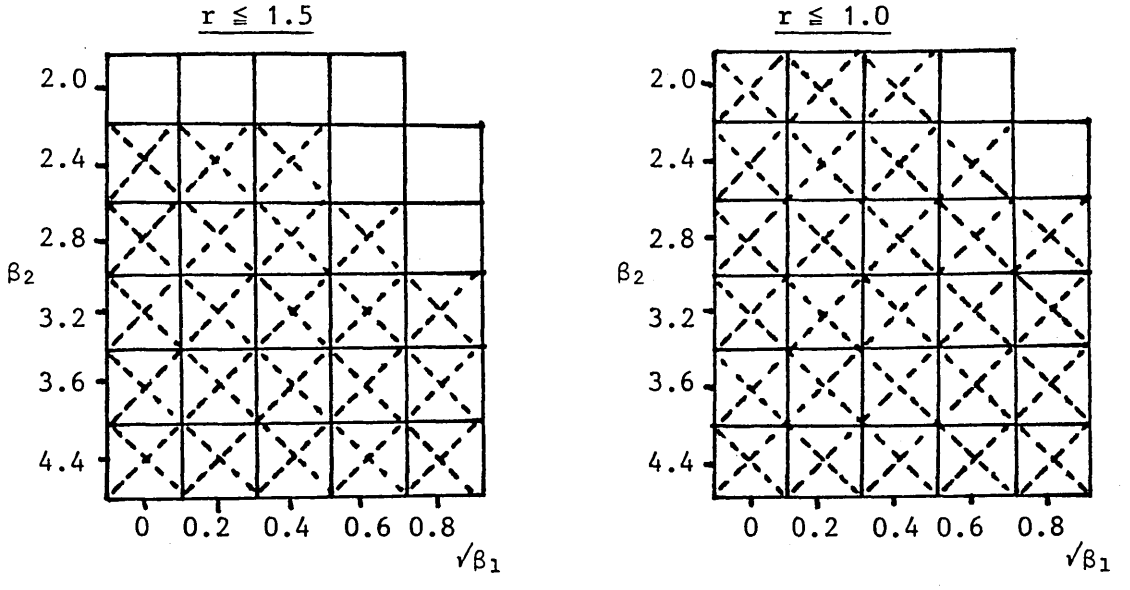
by non-normality. The simulated values of P_J confirmed this.

As with the two sample t-test, under H_1 rounding can cause the differences between the population means μ_i to change. This will result in a change in the alternative hypothesis. When the departure from normality is extreme, rounding will cause the change in H_1 to be greater. The more extreme distributions occur in the top right hand corner of the (β_1,β_2) plane. It is in this region where the change in H_1 due to rounding will be greatest and thus result in less agreement between the P_J and P_{JR} values.

The simulation values of P_{JR} indicated that the effect of rounding on the power levels is greater than for normal populations. Increasing skewness generally widen the range in the P_{JR} values, this being especially so for $\beta_2 < 3$. As expected the distortion in the P_J values due to rounding was greatest in the top right hand corner of the (β_1,β_2) plane. For the values of k considered the power values P_{JR} were of similar magnitude for both n = 10 and 25, except in the region $\beta_2 \leq 2.4$ and $\sqrt{\beta_1} \geq 0.6$. In this region for $r \geq 1.5$ the distortion in the P_J values caused by rounding was greater for n = 25. Table (C.16) gives a selection of P_{JR} values for k = 3 and n = 10. Except for $(\sqrt{\beta_1},\beta_2)$ = (0.6,2.0) and (0.8,2.4), there was reasonable agreement between the range in P_{JR} and P_R (normal population) values for $r \leq 1.0$. This is illustrated by the comparison of P_{JR} values (Table C.16) and P_R values (Table C.19), where k = 3 and n = 10.

As an illustration Figure (5.3.7) indicates the region in the (β_1,β_2) plane where the power is in excess of 0.25 and 0.6 respectively for P = 0.3 and 0.7 for the one way layouts concerned.

Figure 5.3.7: Region in (β_1,β_2) plane where P_{JR} (·) is in excess of 0.25 and 0.60 respectively for P = 0.3 and P = 0.7 for one way analysis of variance where k = 3, 5 and n = 10, 25, when α = 0.05



For normal populations, rounding had a similar effect on the significance and power levels in the one and two way ANOVA. We would expect the same situation for non-normal populations. To confirm this, α_{JR} and P_{JR} values were obtained for a selected number of (β_1,β_2) values. (Appendix C). For the simulations carried out the significance and power levels were in close agreement for the one and two way ANOVA.

5.4 Test Statistic : Exponential Data

Tricker (1984a) considered the effect of rounding on hypothesis tests where the underlying population is assumed exponential. This paper is important for two reasons. It is the only time in the literature where

(a) the effect of rounding on the significance level of a test is considered

(b) the exact sampling distribution of the test statistic is given for rounded data.

This section reviews the work in Tricker (1984a) which is concerned with the effect of rounding on hypothesis testing.

Let $\underline{X} = (X_1,\ldots,X_n)$ be a random sample of size n from an exponential random variable X with unknown parameter θ, with p.d.f. given by (5.4-1)

$$f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \qquad x \geq 0 \qquad \theta > 0 \qquad (5.4\text{-}1)$$

The X_i is subject to rounding where the rounding lattice has interval of width w

and lattice position c.

Then $S_n' = \sum_{i=1}^{n} X_{iR}$ has the following probability distribution.

(Tricker 1984a).

$$P[S_n' = (m+cn)w] = \begin{cases} (k')^n & \text{for } m = 0 \\ [A_{1n}k'^{n-1}ke^{-rc} + A_{2n}k'^{n-2}k^2e^{-2rc(m-1)} \ldots & \\ \qquad A_{nn}k^n e^{-nrc(m-1)} \ldots (m-n+1)]e^{-rm} & \\ & \text{for } m = 1,2,\ldots \end{cases} \qquad (5.4\text{-}2)$$

where $k' = 1 - e^{-r(\frac{1}{2}+c)}$, $k = e^{r/2} - e^{-r/2}$, $A_{jn} = {}^nC_j/(j-1)!$, $r = w/\theta$

A more elegant expression for (5.4–2), which is not given in Tricker (1984a), is:

$$P[S_n' = (m+cn)w] = \begin{cases} (k')^n & \text{for } m = 0 \\ e^{-rm}\left[nk'^{n-1}k + \sum_{\substack{j=2 \\ j\leq m}}^{n} k'^{n-1}k^j e^{-jrc}\, {}^nC_j\, {}^{m-1}C_{j-1}\right] & \text{for } m = 1,2,\ldots \end{cases}$$

For testing the hypothesis $H_0: \theta = \theta_0$ the test statistic is given by

$$T = S_n \quad \text{where } T \sim \frac{\theta_0}{2}\chi^2_{2n} \quad \text{and} \quad S_n = \sum_{i=1}^{n} X_i$$

under rounding

$$T_R = S_n' \quad \text{where the distribution of } T_R \text{ is given by (5.4–2)}$$

For combinations of the values of r, c and n, Tricker compared the distribution of T and T_R. He showed that it is r not c or n which is dominant in determining how close the fit is between T and T_R. For a given rounding lattice and sample size the fit is always better in the right hand tail than in the left hand tail. Finally Tricker showed the effect of rounding on the significance level of the hypothesis test $H_0:\theta = \theta_0$ vs $H_1:\theta < \theta_0$. His results indicate that the significance level can be severely distorted for for $r \geq 1.0$. Although one must be cautious about making statements on such a limited set of results, it appears that the distortion caused by rounding in the significance levels is far greater for tests concerned with exponential populations than for normal populations.

5.5 Discussion and Conclusions

Chapters 3 and 4 recommended ranges of r in which the significance and power level of certain normal tests may be considered acceptable, when the parent population is normal. This chapter has investigated the effect of rounding on these tests when the parent population is non-normal. The main emphasis of this chapter has been to give guidance on what happens to the significance and power level of these tests, if the values of r which were recommended for normal populations are applied when the population is non-normal. Although the study is restricted to $\alpha = 0.05$ and sample sizes of n = 10 and 25, it indicates clearly the behaviour of normal test statistics when applied to rounded non-normal data.

Of all the tests considered the two sample t-test and F-test in the analysis of variance were the most insensitive to rounded non-normal data. In general the results showed that for sample sizes n = 10 and 25, if the value of r recommended for normal populations is used ($r < 2.0$) rounding has very little

effect on the significance level for the entire (β_1,β_2) plane considered. However the power results suggest that, to obtain an acceptable level of power over a large area of the (β_1,β_2) plane, a more suitable level of precision should be $r \leq 1.5$.

For the one sample t and F-tests, departures from normality in the parent population increased the distorting effect of rounding on the significance level of these tests. This distortion was greatest for very skewed distributions found in the top right hand corner of the (β_1,β_2) plane. For both tests, the region in the (β_1,β_2) plane where the significance level is between 3%-7% was found to be very similar for rounded and unrounded data, if the range of r recommended for normal populations is used. For this range of r the level of power within this 3%-7% region of the (β_1,β_2) plane was of a reasonable magnitude, the only exception being for the F-test when $\beta_2 = 2.4$ for $\sqrt{\beta_1} = 0.4$ and 0.6.

In the chi-squared test if the range of r recommended for a normal population ($r \leq 0.5$) is used, the significance and power levels were in fairly close agreement for unrounded and rounded non-normal data, when n = 10 and 25. The only exception to this was for $\beta_2 \leq 2.4$ and $\sqrt{\beta_1} \geq 0.6$. The region in the (β_1,β_2) plane where the significance level lies between 3%-7% was found to be very similar for unrounded and rounded data when n = 10. Increasing the sample size to n = 25 resulted in this similarity decreasing.

In Chapter 5 we have aimed to show how much 'non-normality' can be allowed without the effect of rounding seriously distorting the significance level and power of a test. The evidence suggests that if the range of r recommended for a normal population is used, then these tests can be applied to a wide range of non-normal populations.

This study has considered sample sizes upto n = 25. What about larger sample sizes? We must remember that increasing the sample size does not necessarily diminish the effect of rounding. As an illustration, consider the two sample t–test. When the departure from normality is extreme, rounding can cause the difference between the population means under H_1 to change. This will distort the power of the test. Our results suggest that as n increases, this distortion will become greater.

CHAPTER 6

ESTIMATION OF μ AND σ^2 FOR NORMAL ROUNDED DATA

6.1 Introduction

In the previous chapters we have considered the effects of rounding on test statistics. In this chapter and the next, we shall compare five different procedures for estimating the parameters of a distribution which has been subject to rounding. The first three, maximum likelihood (ML), approximate ML and Sheppard's correction try to compensate for rounding effect. The remaining two, where the method of moments and ML are applied to the midpoint of the rounding intervals will be called naive methods, and make no attempt to compensate.

In this chapter we discuss the five methods of estimation for a normal r.v. X with unknown mean and standard deviation, where X has been rounded. We shall mainly consider these estimation procedures for the degree of precision r up to 2. For larger values of r we are dealing with extremely coarse rounding which is generally unreasonable.

As pointed out earlier in the literature review, work has been carried out by Gjeddebaek (1949, 1956) and Kulldorf (1961) into the properties of maximum likelihood estimates (MLE) of the parameters of grouped normal data. Gjeddebaek (1957, 1959) compared the efficiency of Sheppard's correction with that of ML. He showed that Sheppard's correction is practically as efficient as ML for $r \leq 2.0$ and $n \leq 100$. In section (6.2) it is shown how Gjeddebaek results can be more easily obtained using the results of Chapter 2 and his recommendation for Sheppard's correction is clarified. In the same section the efficiencies of the naive methods of estimation are also examined.

In the past literature several algorithms have been put forward by various authors for finding the MLE of parameters of a distribution, where the data has been rounded. Dempster, Laird and Rubin (1977) pointed out that the EM algorithm may be a useful method. Recently Schader and Schmid (1984) obtained a closed expression for the EM algorithm for grouped normal data. Utilising the results of Chapter 2 an approximate EM–algorithm in section (6.4) is presented for the normal distribution, which is computationally simpler than the 'full' EM algorithm. Its rate of convergence is compared with that of a standard method and is shown to converge slowly for very coarse rounding.

6.2 Maximum Likelihood Estimation (MLE)

Maximum Likelihood estimation of μ and σ^2 in the normal distribution has been covered by such authors as Gjeddebaek (1949, 1956) and Kulldorf (1961), details of which are given in the literature review. A major result of Gjeddebaek's work was obtaining large sample properties of MLE for normal rounded data. He defined the efficiency of these estimates, as the MSE of the MLE from ungrouped data divided by the MSE of MLE for grouped data. Essentially this efficiency is a measure of the loss of information caused by the rounding process. In Gjeddebaek (1957) he showed that the efficiency of the ML estimator of the mean is virtually independent of sample size for $r < 2.0$. When investigating the sample distribution of Sheppard estimators, Gjeddebaek (1959, pp437) assumed the efficiency of the ML estimator of the variance to be approximately the same for all sample sizes. Results from a simulation study in section (6.4) confirm this.

6.3 Other Methods Of Estimation

As shown by Tallis (1967) for the normal distribution when the data has been rounded, the approximate ML and Sheppard's method produce the same estimates of mean and variance.

$\underline{X} = (X_1 \ldots X_n)$ is a random sample of size n from a normal random variable X with mean μ and variance σ^2. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the rounded value of X_i corresponding to a rounding lattice with interval of width w and lattice position c. The estimates of μ and σ^2 by Sheppard's method are:

$$\tilde{\mu}_R = \bar{X}_R \qquad \tilde{\sigma}_R^2 = S_R^2 - \frac{w^2}{12} \qquad (6.3\text{-}1)$$

where

$$\bar{X}_R = \sum_i \frac{X_{Ri}}{n} \quad , \quad S_R^2 = \frac{1}{n-1} \sum_i (X_{Ri} - \bar{X}_R)^2$$

(6.3-1) are also the estimators given by the approximate ML method. From now on we shall refer to them simply as the Sheppard estimators for μ and σ^2.

For $r \leq 2.0$ and $n \leq 100$ Gjeddebaek (1959) states that Sheppard estimators are almost as efficient as ML estimators. This statement has often been misinterpreted. For example, Krusal and Tanur (1978) states that Sheppard estimators of μ and σ^2 have the same efficiency as the ML estimators, for $r \leq 2.0$ and $n \leq 100$. This is not strictly correct. In order to clarify Gjeddebaek's statement we proceed as follows.

To express how 'good' Sheppard estimators are, Gjeddebaek compared the MSE of these estimators for rounded and unrounded data. He defined the efficiency of a Sheppard estimator as the MSE of this estimator for unrounded data divided by the MSE of the estimator in question for rounded data. This gives

$$\epsilon(\tilde{\mu}_R, \tilde{\mu}) = \frac{MSE(\tilde{\mu})}{MSE(\tilde{\mu}_R)} \qquad \epsilon(\tilde{\sigma}_R^2, \tilde{\sigma}^2) = \frac{MSE(\tilde{\sigma}^2)}{MSE(\tilde{\sigma}_R^2)} \qquad (6.3\text{-}2)$$

where

$$\tilde{\mu} = \bar{X} \text{ and } \tilde{\sigma}^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

Essentially the efficiencies in (6.3-2) are a measure of the loss of information caused by the rounding process.

Using the same definition of efficiency as Gjeddebaek, we require the MSEs of Sheppard estimators. By using the results of Chapter 2, these can be obtained by a different method from Gjeddebaek's.

The mean and variance of $\tilde{\mu}_R$ and $\tilde{\sigma}^2_R$ are given by:

$$E[\tilde{\mu}_R] = E[\bar{X}_R] = \mu_R \qquad E[\tilde{\sigma}_R^2] = E[S_R^2] - \frac{w^2}{12} = \sigma_R^2 - \frac{w^2}{12}$$

$$V[\tilde{\mu}_R] = V[\bar{X}_R] = \frac{\sigma_R^2}{n} \qquad V[\tilde{\sigma}_R^2] = V\left[S_R^2 - \frac{w^2}{12}\right]$$

$$= \left[\frac{2}{n-1} + \frac{\beta_{2R} - 3}{n}\right]\sigma_R^4$$

from (3.3-10)

hence MSEs:

$$MSE[\tilde{\mu}_R] = \frac{\sigma_R^2}{n} + (\mu_R - \mu)^2 \qquad (6.3\text{-}3)$$

$$MSE[\tilde{\sigma}_R^2] = \left[\frac{2}{n-1} + \frac{\beta_{2R} - 3}{n}\right]\sigma_R^4 + \left[\sigma_R^2 - \frac{w^2}{12} - \sigma^2\right]^2$$

As in Gjeddebaek (1957, 1959), we shall assume a normal distribution with mean zero and variance one. By using a standard normal distribution we lose no generality in the results.

The expressions in (6.3-3) are computationally simpler than those given in Gjeddebaek (1957, 1959). They only require the calculation of μ_R, σ^2_R and β_{2R}. These can be evaluated from expressions in (2.2-16) Chapter 2. These expressions quickly converge and are easy to calculate. However Gjeddebaek's expressions for MSE require the calculation of the normal distribution function several times and are troublesome to handle.

Using the expression in (6.3-3) and the MSE for the Sheppard estimators $(\tilde{\mu},\tilde{\sigma})$ for unrounded data, the efficiencies of Sheppard estimators were calculated. Table (6.3.1) presents the results for $\epsilon(\tilde{\mu}_R,\tilde{\mu})$ and $\epsilon(\tilde{\sigma}^2_R,\tilde{\sigma}^2)$ for r = 2.0 and 1.5, for three lattice positions. Gjeddebaek in his paper has given a smaller set of results which are in agreement with ours. In the table the efficiencies of Sheppard estimators are compared with those for ML estimators for rounded data when n is large given in Gjeddebaek (1957). As stated in section (6.2), this is reasonable as the efficiencies for ML estimators are approximately the same for all n for $r \leq 2.0$. From the results in Table (6.3.1) the following may be stated:

For $r \leq 1.5$ Sheppard and ML estimators have the same loss in efficiency when applied to rounded data.

For $1.5 \leq r \leq 2.0$ and $n \leq 100$ the loss in efficiency in Sheppard estimators can be up to 15% more than ML estimators.

In Gjeddebaek (1959) it stated that Sheppard estimators of μ and σ^2 are practically as efficient as ML estimators. The above is a clear statement of what is meant by practically as efficient. It removes the confusion that it implies that Sheppard estimators have the same efficiency as ML estimators for $n \leq 100$ and $r \leq 2.0$.

TABLE 6.3.1 : Efficiency Of Estimates Obtained By Sheppard's Method

		Lattice Position					
n		c = 0		c = 0.25		c = 0.5	
		$\epsilon(\tilde{\mu}_R,\tilde{\mu})$	$\epsilon(\tilde{\sigma}^2_R,\tilde{\sigma}^2)$	$\epsilon(\tilde{\mu}_R,\tilde{\mu})$	$\epsilon(\tilde{\sigma}^2_R,\tilde{\sigma}^2)$	$\epsilon(\tilde{\mu}_R,\tilde{\mu})$	$\epsilon(\tilde{\sigma}^2_R,\tilde{\sigma}^2)$
r = 1.5	10	84.3	71.4	84.2	72.0	84.2	72.6
	100	84.3	71.3	84.2	72.0	84.2	72.6
	100,000	84.3	71.3	84.2	71.9	84.2	71.5
r = 1.5	MLE large n	84.6	72.2	84.2	72.1	84.7	71.8
r = 2.0	10	74.8	50.3	74.8	58.2	73.2	68.7
	100	74.8	48.5	74.4	58.4	73.2	68.5
	500	74.8	43.9	74.4	58.4	73.2	60.2
	1000	74.8	39.8	72.7	58.4	73.2	52.5
r = 2.0	MLE large n	75.5	63.5	75.1	58.7	74.6	54.1

Although Gjeddebaek has shown that Sheppard and ML estimators have a similar loss in efficiency when applied to rounded data; we need to compare these methods of estimation directly under rounding. In order to compare Sheppard and ML methods, the efficiency of Sheppard estimators relative to ML estimators will be defined as follows:

$$\epsilon(\tilde{\mu}_R, \hat{\mu}_R) = \frac{MSE(\hat{\mu}_R)}{MSE(\tilde{\mu}_R)} \quad , \quad \epsilon(\tilde{\sigma}^2_R, \hat{\sigma}^2_R) = \frac{MSE(\hat{\sigma}^2_R)}{MSE(\tilde{\mu}_R)}$$

where $(\tilde{\mu}_R, \tilde{\sigma}^2_R)$ and $(\hat{\mu}_R, \hat{\sigma}^2_R)$ are respectively the Sheppard and ML estimators for rounded data.

Table (6.3.2) presents these efficiencies for r = 2.0, r = 1.5 and r = 0 (no rounding) for three lattice positions.

TABLE (6.3.2) : Efficiency of Sheppard Estimators Relative to ML Estimators

	n	Lattice Position					
		c = 0		c = 0.25		c = 0.5	
		$\epsilon(\tilde{\mu}_R, \hat{\mu}_R)$	$\epsilon(\tilde{\sigma}^2_R, \hat{\sigma}^2_R)$	$\epsilon(\tilde{\mu}_R, \hat{\mu}_R)$	$\epsilon(\tilde{\sigma}^2_R, \hat{\sigma}^2_R)$	$\epsilon(\tilde{\mu}_R, \hat{\mu}_R)$	$\epsilon(\tilde{\sigma}^2_R, \hat{\sigma}^2_R)$
r = 1.5	10	99.6	84.7	100	85.4	99.4	86.4
	20	99.6	91.9	100	92.9	99.4	94.1
	100	99.6	97.1	100	98.4	99.4	99.6
	1000	99.6	98.0	100	99.1	99.4	99.6
r = 2	10	99.1	71.0	99.6	84.8	98.1	108.6
	20	99.1	74.1	99.2	92.6	98.1	118.1
	100	99.1	76.2	99.0	99.5	98.1	124.7
	1000	99.1	62.0	96.0	99.5	98.1	97.0
	10,000	99.1	22.5	86.2	99.5	98.1	28.8

	n	$\epsilon(\tilde{\mu}_R, \hat{\mu}_R)$	$\epsilon(\tilde{\sigma}^2_R, \hat{\sigma}^2_R)$
r = 0	10	100	85.5
	20	100	93.1
	100	100	98.5
	1000	100	99.9

The results from Table (6.3.2) indicate that for $r \leq 1.5$, Sheppard estimators are almost as efficient as ML estimators. For this degree of precision there is little differnce between the efficiencies for rounded and unrounded data. For r = 2.0, this is not the case. The efficiency $\epsilon(\tilde{\sigma}^2_R, \hat{\sigma}^2_R)$ can be considerably lower as n increase in size. However, for $r \leq 2.0$ and $n \leq 100$ it may be of little preference to use ML method instead of Sheppard's method. The reduction in efficiency in using Sheppard estimators may be worth while in view of their computational simplicity.

Naive Methods

Often rounding of data is ignored and the midpoint of the rounding intervals are used to estimate the parameters of a distribution. We briefly investigate this naive approach using the method of moments and ML. For the normal distribution they result in the same estimators of μ and σ^2

$$\hat{\mu}_N = \bar{X}_R \quad , \quad \hat{\sigma}^2_N = \frac{1}{n}\sum_i (X_{Ri}-\bar{X}_R)^2 = \left[1 - \frac{1}{n}\right]S_R^2 \qquad (6.3\text{-}4)$$

In this section we investigate the efficiency of the naive estimators $(\hat{\mu}_N,\hat{\sigma}^2_N)$ relative to Sheppard estimators $(\tilde{\mu}_R,\tilde{\sigma}^2_R)$. The naive estimator of μ is the same as that given by Sheppard's method and needs no further investigation.

The mean and variance of $\hat{\sigma}^2_N$ are given by

$$E[\hat{\sigma}^2_N] = E\left[\left[1 - \frac{1}{n}\right]S_R^2\right] = \left[1 - \frac{1}{n}\right]\sigma_R^2$$

$$V[\hat{\sigma}^2_N] = V\left[\left[1 - \frac{1}{n}\right]S_R^2\right] = \left[1 - \frac{1}{n}\right]^2 V[S_R^2] \qquad \text{from (6.3-3)}$$

$$= \left[1 - \frac{1}{n}\right]^2\left[\frac{2}{n-1} + \frac{\beta_{2R}-3}{n}\right]\sigma_R^4$$

hence MSE is

$$MSE(\hat{\sigma}^2_N) = \left[1 - \frac{1}{n}\right]^2\left[\frac{2}{n-1} + \frac{\beta_{2R}-3}{n}\right]\sigma_R^4 + \left[\left[1 - \frac{1}{n}\right]\sigma_R^2 - \sigma^2\right]^2 \qquad (6.3\text{-}5)$$

Although the variance of $\hat{\sigma}^2_N$ is smaller than $\tilde{\sigma}^2_R$, the bias in $\hat{\sigma}^2_N$ is generally larger. Because of the bias in $\hat{\sigma}^2_N$, the MSE of this estimator can be considerably larger than for the Sheppard estimator $\tilde{\sigma}^2_R$ (6.3–3). For certain values of n and r the naive estimator could be at least as efficient in terms of MSEs as the Sheppard estimator. For example, this was true for $r < 1.0$ where $n < 50$ and $r < 1.5$ where $n < 20$.

6.4 Approximate Expectation – Maximisation (EM) Algorithm

Several authors considered the problem of finding MLE of μ and σ^2 for rounded data, for example Gjeddebaek (1949), Kulldorf (1961) and Swann (1969). More recently Dempster, Laird and Rubin (1977) proposed that the Expectation–Maximisation (EM) algorithm as a suitable method for obtaining the MLE for grouped data. Schader and Schmid (1984) were the first to present an expression for the EM algorithm for data from a normal distribution subject to rounding. In this section using the results of Chapter 2 an approximate EM algorithm is presented which is computationally simpler than that given by Schader and Schmid (1984). Because of its reliability, it makes it an attractive alternative to other methods. For completeness the EM method is summarised below.

For a given parameterisation $\underline{\theta}$ choose a complete data sufficient statistic $t = t(\underline{X})$ ie a statistic sufficient for $\underline{\theta}$ when $\underline{X}$ is full observation. Starting at estimate $\underline{\theta}_P$ compute the conditional expectation of $\underline{t}_P$ of $\underline{t}$ given the rounded data and assuming $\underline{\theta}_P$ to be true. This is known as the E–step. Next compute $\underline{\theta}_{P+1} = \hat{\theta}(\underline{t}_P)$ where $\hat{\underline{\theta}}(\underline{t})$ is the MLE of $\underline{\theta}$ when complete data sufficient statistic $\underline{t}$ has been observed, this is the M–step.

The approximate EM algorithm can be obtained in the following way:

$\underline{X} = (X_1 \ldots X_n)$ is a random sample of size n from a normal random variable X with mean μ and variance σ^2 with p.d.f. $f(x \mid \mu, \sigma^2)$. The X_i are subject to rounding when the rounding lattice has intervals of width w and lattice position c. Let n_j be the number of sample values with midpoint $Y_j = cw + jw$, for some integral value of j, where $\Sigma n_j = n$. The lower and upper boundaries of each rounding interval are given by $L_j = Y_j - {}^{w}/_{2}$ and $U_j = Y_j + {}^{w}/_{2}$.

The joint sufficient statistics for μ and σ are

$$t_1(\underline{X}) = \sum_i X_i \quad \text{and} \quad t_2(\underline{X}) = \sum_i X_i^2$$

Let μ_P, σ^2_P be the pth estimate of μ^2 and σ^2. The E–step is given by:

$$E\left[\sum_i X_i \mid \underline{Y}, \underline{n}, \mu_P, \sigma_P^2\right] = n \sum_j \left[\frac{n_j}{n}\right] \left[\frac{\int_{L_j}^{U_j} x f(x \mid \mu_P, \sigma_P^2)\,dx}{\int_{L_j}^{U_j} f(x \mid \mu_P, \sigma_P^2)\,dx}\right] \tag{6.4-1}$$

$$E\left[\sum_i X_i^2 \mid \underline{Y}, \underline{n}, \mu_P, \sigma_P^2\right] = n \sum_j \left[\frac{n_j}{n}\right] \left[\frac{\int_{L_j}^{U_j} x^2 f(x \mid \mu_P, \sigma_P^2)\,dx}{\int_{L_j}^{U_j} f(x \mid \mu_P, \sigma_P^2)\,dx}\right]$$

where

$$\underline{Y} = (\ldots Y_{-1}, Y_0, Y_1 \ldots) \quad \text{and} \quad \underline{n} = (\ldots n_{-1}, n_0, n_1 \ldots)$$

If the sample size is reasonably large and rounding coarse, then to the first order of approximation

$$\left[\frac{n_j}{n}\right] = \int_{L_j}^{U_j} f(x|\mu_P,\sigma_P^2)dx$$

From (6.4-1) it follows that an approximate E step is

$$E\left[\sum_i X_i \middle| \underline{Y},\ \underline{n},\ \mu_P,\ \sigma_P^2\right] = n \sum_J \int_{L_j}^{U_j} x\ f(x|\mu_P,\sigma_P^2) = n\ E[X|\mu_P,\sigma_P^2]$$

$$E\left[\sum_i X_i \middle| \underline{Y},\ \underline{n},\ \mu_P,\ \sigma_P^2\right] = n \sum_J \int_{L_j}^{U_j} x^2 f(x|\mu_P,\sigma_P^2) = n\ E[X^2|\mu_P,\sigma_P^2]$$

The MLEs for unrounded data are

$$\sum_i \frac{X_i}{n} \quad \text{and} \quad \frac{\Sigma X_i^2}{n} - \left[\frac{\Sigma X_i}{n}\right]^2$$

hence the M step is

$$\begin{aligned} \mu_{P+1} &= E[X|\mu_P,\sigma_P^2] \\ \sigma_{P+1}^2 &= V[X|\mu_P,\sigma_P^2] \end{aligned} \qquad (6.4\text{-}2)$$

Using (2.2-13) and (2.2-15) and expressions for the mean and variance of X are given by:

$$\begin{aligned} E[X|\mu_P,\sigma_P^2] &= E[X_R] - E_1(\mu_P,\sigma_P^2] \\ V[X|\mu_P,\sigma_P^2] &= V[X_R] - \frac{w^2}{12} - E_2(\mu_P,\sigma_P^2] \end{aligned} \qquad (6.4\text{-}3)$$

where

$$E_1(\mu_P,\sigma_P^2] = \frac{w}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} D \, \mathrm{Sin}\left[2\pi k\left[\frac{\mu_P}{w} - c\right]\right]$$

$$E_2(\mu_P,\sigma_P^2) = 4 \sum_{k=1}^{\infty} (-1)^k \left[\sigma_P^2 + \left[\frac{w}{2\pi k}\right]^2\right] D \, \mathrm{Cos}\left[2\pi k\left[\frac{\mu_P}{w} - c\right]\right] - \left[E_1(\mu_P,\sigma_P^2)\right]^2$$

where

$$D = \exp\left[-\, 2\pi^2 k^2 \sigma_P^2 / w^2\right]$$

If n is reasonably large $\bar{X}_R$ and S^2_R will be precise estimates of $E[X_R]$ and $V[X_R]$ respectively. Thus (6.4-3) may be approximated by

$$\begin{aligned} E[X|\mu_P,\sigma_P^2] &= \bar{X}_R - E_1(\mu_P,\sigma_P^2) \\ V[X|\mu_P,\sigma_P^2) &= S_R^2 - \frac{w^2}{12} - E_2(\mu_P,\sigma_P^2) \end{aligned} \qquad (6.4\text{-}4)$$

Substituting (6.4-4) into (6.4-2) gives the approximate EM algorithm as:

$$\begin{aligned} \mu_{P+1} &= \bar{X}_R - E_1(\mu_P,\sigma_P^2) \\ \sigma^2_{P+1} &= S_R^2 - \frac{w^2}{12} - E_2(\mu_P,\sigma_P^2) \end{aligned} \qquad (6.4\text{-}5)$$

The approximate EM has two advantages over the 'full' EM given by Schader and Schmid (1984). The approximate EM is computationally simpler and does not require the evaluation of normal density and distribution functions. This fact becomes more important as n increase, as the number of cells containing

observations will also increase. This will result in the normal routines which calculate the density and distribution functions being called more often. Secondly the approximate EM requires less information from the sample. The full EM algorithm requires the number of observations in each cell (n_j) whereas the approximation does not.

In obtaining the approximate EM it was assumed that the rounding was coarse and sample size large. Under such conditions the number of rounding intervals will be few and n_j/n should be a good approximation to the probability of an observation falling in a particular interval. Also $\bar{X}_R$ and S^2_R should be precise estimates of $E[X_R]$ and $V[X_R]$ respectively. As a result we would expect the approximate EM to produce accurate MLE when the rounding is coarse and sample size large. To ascertain the region of (n,r) where the approximate EM is a reliable method for obtaining the MLE, we proceed as follows.

To test the approximate EM algorithm we must compare its results with those of the 'true' MLE. Any standard method which gives us the MLE would have been adequate. Also worthwhile is to investigate the rate of convergence of the approximate EM algorithm.

Schader and Schmid (1984) adapted an algorithm used by Van-Wärden (1973) by using Taylor Series to obtain the MLE for grouped normal data. They showed that this algorithm is identical with the algorithm obtained with Fisher's method of scoring, but does not require any second derivatives. It was demonstrated that this algorithm SCOR gave a better average number of iterations than did Newton-Raphson and Fixed Point methods. We decided to use SCOR to obtain the 'true' MLE as it also provided information on how the number of iterations of

the approximate EM compared with that for a very efficient algorithm.

The MLEs produced by the approximate EM and SCOR algorithms were compared by simulation study, as were their rates of convergence. The simulation study used a purpose-written program, with Nag routines used to generate random samples of normal deviates. The normal deviates were rounded before MLEs were obtained, using both algorithms. The mean and variance of the normal deviates were set to 0 and 1 respectively. By using a standard normal distribution we lose no generality in the results. Using a normal distribution with mean μ and variance σ^2 would simply cause the difference between the MLE of μ and σ^2 produced by the two algorithms to be multiplied by σ and σ^2 respectively. In the study the rounding precision r varied in the range 0.5 to 3.0 for lattice positions c = 0.0, 0.25 and 0.5. Sample sizes considered were $n \geq 50$.

For each value of r and c the MLEs were obtained 1000 times for each algorithm. From these 1000 replicates estimates of the mean and standard error of the MLE of μ and σ^2 were obtained, together with the average number of iterations taken for each algorithm to converge. Convergence to MLE had occurred if the absolute values of the differences between successive estimates of μ and σ^2 were both less than 10^{-5}. In all runs considered both algorithms converged for the same samples. In some cases to compare the results given by the two algorithms in detail, the actual MLEs were produced. To check the validity of the program a sample of size 2000 was run. The MSEs of the estimators of μ and σ^2 were in close agreement with the theoretical results given by Gjeddebaek (1956).

For rounded data there is a probability that the MLEs of μ and σ^2 do not exist. This probability will tend to zero as the sample size increases. Essentially we are dealing with conditional MLE, the condition being existence. Kulldorf (1961) gives sufficient conditions for an MLE to exist. In the present study their existence is not a problem as the sample sizes are large. Although we are dealing with conditional MLEs we shall refer to them merely as the MLEs.

Simulation Results

The means and standard errors of the MLEs of μ and σ^2 for the approximate EM and SCOR methods were in close agreement in the region $r \geq 1.0$ and $n \geq 100$. This result was not surprising as the MLEs for individual samples were also in close agreement for the same region. Table (6.4.1) shows the results for $r = 1.0$, $c = 0$, $n = 100$ and $r = 3$, $c = 0$ and $n = 500$. As expected these results in the table show that as r and n increase the difference between MLEs from the two methods decreases. To obtain the differences between the MLEs produced by the two algorithms, when the distribution is normal with mean μ and variance σ^2, we simply multiply columns 1 and 2 of Table (6.4.1) by σ and σ^2 respectively.

By obtaining MLEs from the SCOR algorithm for $n \geq 10$ we confirmed Gjeddebaek's (1959) assumption concerning the ML estimator of σ^2. The results endorse his conjecture that the ML estimator of σ^2 has approximately the same efficiency for all sample sizes, where efficiency is defined as in section (6.2).

TABLE 6.4.1a : Sample of MLE From Approximate EM and SCOR Algorithms

r = 1.0 c = 0 n = 100				r = 3.0 c = 0 n = 500			
EM		SCOR		EM		SCOR	
$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{\sigma}^2$
-0.120	0.882	-0.120	0.881	0.000	1.096	0.000	1.096
-0.010	0.767	-0.010	0.766	0.135	0.873	0.135	0.873
0.240	1.100	0.240	1.100	-0.008	0.961	-0.008	0.961
0.070	0.802	0.070	0.801	0.029	1.095	0.029	1.095
0.000	1.217	0.000	1.216	0.110	0.869	0.110	0.869
-0.040	0.815	-0.040	0.816	-0.008	1.023	-0.008	1.022
-0.020	0.976	-0.020	0.976	0.007	1.150	0.007	1.150
0.080	1.270	0.080	1.270	0.086	0.974	0.086	0.974
0.200	0.877	0.200	0.875	-0.016	0.992	-0.016	0.992
0.240	0.699	0.241	0.697	-0.024	0.941	-0.024	0.941

TABLE 6.4.1b : Summary Statistics for MLE

	Mean (Standard Error)			
	EM		SCOR	
	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{\sigma}^2$
r = 1.0, c = 0 n = 100	0.002(0.104)	0.991(0.150)	0.002(0.105)	0.990(0.149)
r = 3.0, c = 0 n = 500	0.001(0.063)	1.000(0.077)	0.001(0.063)	1.000(0.077)

Dempster, Laird and Rubin (1977) stated that the linear rate of convergence of the EM algorithm is dependent on the amount of information lost: the smaller the loss in information, the quicker the rate of convergence. The results for the approximate EM clearly display this. For example, for a sample of size 100 with r = 1.5 and c = 0, the average number of iterations for convergence was 3.1. This compared with 5.5 when r = 2.0 and c = 0. Table (6.4.2) shows the range for the average number of iterations taken for each algorithm to converge when n = 250, over the three lattice positions considered.

TABLE 6.4.2

	Average number of iterations when n = 250 for c = 0, 0.25 and 0.5	
r	Approx EM	SCOR
1.0 - 2.0	2 - 5.6	3 - 4.1
2.5	7.5 - 13.4	3.5 - 4.5
3.0	17.1 - 43.5	3.4 - 6.4

Although Table (6.4.2) shows only the number of iterations for n = 250, they do represent closely the results for other sample sizes.

6.5 Conclusions

In this chapter the estimation of μ and σ^2 for normal rounded data where $r \leq 2.0$ has been considered. Using the results of Chapter 2, it has been established that for $r \leq 1.5$ there is little difference between the efficiency of Sheppard's and ML methods for estimating μ and σ^2. Even for $r \leq 2.0$ and $n \leq 100$ it may be of little preference to use the ML method. The loss in efficiency in using Sheppard

estimators is far outweighed by their computational simplicity. For values of r up to 2, the ML method is preferrable in terms of efficiency over Sheppard's method, in the region $r = 2$ and $n > 100$. In this region of r and n the approximate EM gave reliable results with a rate of convergence only slightly less than that for the SCOR method. Together with its simplicity is is an attractive alternative to other standard methods for obtaining MLEs.

Generally the approximate EM was found to be a suitable method for obtaining MLEs for $r \geq 1.0$ and $n \geq 100$. However its rate of convergence was slow for $r > 2.0$. This should be borne in mind whenever an EM approach is considered. This is something which has not been demonstrated in the literature before for rounded data.

The naive estimators were found to be at least as efficient as Sheppard estimators for only a restricted range of r and n. Hence Sheppard estimators are preferrable as they generally have a higher efficiency and are computationally as simple.

CHAPTER 7

ESTIMATION OF PARAMETERS FOR ROUNDED DATA FROM NON-NORMAL DISTRIBUTIONS

7.1 Introduction

Studies of estimation from rounded data have concentrated largely on the normal distribution. For this distribution, estimation procedures have been found to be very robust to the effects of rounding data. Chapter 6 showed that the ML and Sheppard estimators of μ and σ^2 have a high efficiency even for coarsely rounded data. In this chapter we investigate if this is true for other distributions too.

The past literature contains very little on the estimation of parameters in non-normal distributions, subject to rounding. The MLE of the exponential was given by Kulldorff (1961) for grouped data. Tricker (1984a) considered the estimation of the exponential parameter in terms of MSE. Tallis and Young (1962) obtained the MLE of the parameters of the log normal for grouped data. In these papers or in previous literature, there has been no work concerning the efficiency of various estimation procedures, applied to non-normal rounded data. Generally only the ML method has been applied to non-normal rounded data. Other methods of estimation have not been considered.

There are many non-normal distributions that could be looked at. In this chapter two such distributions have been chosen to demonstrate the implications of rounding. The exponential has been chosen because of its simplicity and the gamma as this is a more complicated p.d.f., where the shape of the distribution can range from near normality to extreme non-normality. As in Chapter 6 the same five estimation procedures will be discussed for these two distributions for r upto 2. To find the MLE for gamma rounded data, the EM algorithm is used. This is the first time this algorithm has been applied to non-normal rounded data.

7.2 Gamma Distribution

In this section the five methods of estimation are discussed for a gamma random variable X, with unknown parameters α and θ, where X has been rounded. The p.d.f. for the gamma distribution is given by

$$f(x|\theta,\alpha) = \frac{1}{\Gamma(\alpha)} \frac{1}{\theta} \left[\frac{x}{\theta}\right]^{\alpha-1} e^{-\frac{x}{\theta}} \qquad \begin{matrix} x \geq 0 \\ \theta,\alpha > 0 \end{matrix} \qquad (7.2\text{-}1)$$

As far as the author is aware the estimation of the parameters for gamma rounded data has not appeared in previous literature.

For gamma rounded data explicit expressions for the MLE of the parameters are unavailable, hence the appropriate distributions cannot be examined directly. However some insight could be obtained by examining large sample properties of the MLE. For the method of moments and Sheppard's method, explicit expressions can be found for the estimators. However like ML estimators, it is still possible to derive only large sample properties. This study is interested also in moderate size n and these large sample results would at best be only approximate. For ML estimators for unrounded data, the large sample results are rather inaccurate unless n is very large (Lawless 1982). There is no reason to believe that this does not extend to rounded data. Because of the inaccuracy of the large sample properties for moderate n, the five estimation procedures were examined by simulation.

The simulation study used a single purpose written program, with Nag routines used to generate random samples of gamma deviates. These deviates were rounded

before the various estimators were obtained. In addition the ML estimator for unrounded data was calculated. The rounding precision r varied upto 2, for lattice positions c = 0, 0.25 and 0.5. Sample sizes considered were n = 25, 50, 100 and 500. Fixing θ = 1, α took the values 1, 5, 10 and 20. Setting θ = 1 would not alter the generality of the results and the range of α would span various degrees of non-normality. For various combinations of r, c, n and α, each estimator was obtained 2000 times. From these 2000 replicates the mean, variance and MSE of estimators were found. The number of iterations taken for the EM algorithm to converge was recorded. The number of replicates was limited to 2000 due to the complex calculations involved and the iterative nature of the simulation. A larger simulation was not practical on the Polytechnic computer.

Several procedures were used to verify the validity of the program. For example, to check the generation of gamma deviates, the values of α and θ given by the simulation were compared with their expected values.

7.2.1 Maximum Likelihood Estimation

In this section the main aim is to find the MLE for rounded data. Various standard methods could have been used. However, it was decided to use the EM methods because, as already mentioned in section (6.1), this has been put forward as a possible approach for rounded data. Although the EM may be slower to converge than some other methods, this must be put against the fact that it is simple to program and robust. (For example, it is less concerned about initial starting values). The EM algorithm may be obtained in the following way.

EM–Algorithm for Gamma Distribution

$\underline{X} = (X_1 \ldots X_n)$ is a random sample of size n from a gamma random variable X with parameters α and θ with p.d.f. $f(x|\alpha,\theta)$ given by (7.2–1). The X_i are subject to rounding where the rounding lattice has intervals of width w and lattice position c. Let n_j be the number of sample values with midpoints $Y_j = cw + jw$ for some positive integral value j,

where $\sum_j n_j = n.$

The lower and upper boundaries of each rounding interval are given by

$$\ell_j = Y_j - \frac{w}{2} \text{ and } u_j = Y_j + \frac{w}{2}.$$

The joint sufficient statistics for α and θ are

$$\sum_i X_j \text{ and } \prod_i X_i$$

In this problem it is more convenient to deal with the sufficient statistics

$$\sum_i X_i/n \text{ and } \sum_i \log X_i/n.$$

This will not affect the algorithm's convergence behaviour as it is independent of the sufficient statistics used. Let α_p and θ_p be the pth estimates of α and θ respectively. For the E step we require:

$$E\left[\sum_i X_i \,|\, \underline{Y},\ \underline{n}\ \theta_p,\ \alpha_p\right] \qquad (7.2\text{-}2)$$

$$E\left[\sum \log X_i \,|\, \underline{Y},\ \underline{n},\ \theta_p,\ \alpha_p\right] \qquad (7.2\text{-}3)$$

where $\underline{Y} = (Y_0, Y_1, \ldots)$ and $\underline{n} = (n_0, n_1, \ldots)$.

An expression for (7.2–3) can be found as follows:

For $j = 1,2,\ldots$

$$E[X_i \mid Y_j, n_j, \theta_p, \alpha_p] = \frac{\frac{1}{\theta_p \Gamma(\alpha_p)} \int_{\ell_j}^{u_j} x \left[\frac{x}{\theta_p}\right]^{\alpha_p - 1} \exp\left[-\frac{x}{\theta_p}\right] dx}{\frac{1}{\theta_p \Gamma(\alpha_p)} \int_{\ell_j}^{u_j} \left[\frac{x}{\theta_p}\right]^{\alpha_p - 1} \exp\left[-\frac{x}{\theta_p}\right] dx}$$

$$= \alpha_p \theta_p \left[\frac{G[u_j/\theta_p, \alpha_{p+1}] - G[\ell_j/\theta_p, \alpha_{p+1}]}{G[u_j/\theta_p, \alpha_p] - G[\ell_j/\theta_p \alpha_p]}\right]$$

where the incomplete gamma function is given by

$$G[x,p] = \frac{1}{\Gamma(p)} \int_0^x \exp(-t)\, t^{p-1}\, dt$$

For the interval containing zero ie the first interval (j=0)

$$E[X_i \mid Y_0, n_0, \alpha_p, \theta_p] = \alpha_p \theta_p \left[\frac{G[u_0/\theta_p, \alpha_{p+1}]}{G[u_0/\theta_p, \alpha]}\right]$$

Hence (7.2–2) denoted by E_{1p} is

$$E_{1p} = \frac{\alpha_p \theta_p}{n}\left[n_0 \left\{\frac{G[u_0/\theta_p, \alpha_{p+1}]}{G[u_0/\theta_p, \alpha]}\right\} + \sum_{j=1}^{\infty} n_j \left[\frac{G[u_j/\theta_p, \alpha+1] - G[\ell_j/\theta_p, \alpha_{p+1}]}{G[u_j/\theta_p, \alpha] - G[\ell_j/\theta_p, \alpha_p]}\right]\right] \qquad (7.2\text{-}4)$$

To obtain an expression for (7.2-3):

For $j = 1,2,\ldots$

$$E[\log X_i | Y_j, n_j, \alpha_p, \theta_p] = \frac{\int_{\ell_j/\theta_p}^{u_j/\theta_p} \log(x\theta_p) x^{\alpha_p - 1} e^{-x}\, dx}{\Gamma(\alpha_p)\left[G[u_j/\theta_p, \alpha_p] - G[\ell_j/\theta_p, \alpha_p]\right]}$$

For $j = 0$, first interval

$$E[\log X_i | Y_0, n_0, \alpha_p, \theta_p] = \frac{\int_{0}^{u_j/\theta_p} \log(x\theta_p) x^{\alpha_p - 1} e^{-x}\, dx}{\Gamma(\alpha_p) G[u_0/\theta_p, \alpha_p]}$$

Hence (7.2-3) denoted by E_{2p} is given by:

$$E_{2p} = \frac{1}{n\Gamma(\alpha_p)} \left[\frac{n_0 I_0}{G[u_0/\theta_p, \alpha_p]} + \sum_{j=1}^{\infty} \frac{n_j I_j}{G[u_j/\theta_p, \alpha_p] - G[\ell_j/\theta_p, \alpha_p]}\right] \qquad (7.2\text{-}5)$$

where $$I_0 = \int_0^{u_j/\theta_p} \log(x\theta_p) x^{\alpha_p - 1} e^{-x}\, dx$$

$$I_j = \int_{\ell_j/\theta_p}^{u_j/\theta_p} \log(x\theta_p) x^{\alpha_p - 1} e^{-x}\, dx \qquad (j=1,2,\ldots)$$

The M step is given by

$$\log(\alpha_{p+1}) - \psi(\alpha_{p+1}) - \log(E_{1p}) + E_{2p} = 0 \qquad (7.2\text{–}6)$$

$$\theta_{p+1} = E_{1p}/\alpha_{p+1}$$

where $\psi(z) = \dfrac{d \log \Gamma(z)}{dz}$, the digamma function.

Given the starting estimates (α_p, θ_p) the next estimates $(\alpha_{p+1}, \theta_{p+1})$ are obtained by solving (7.2–6) (p=0,1,2,...) in the normal way. (7.2–6) is the EM algorithm for estimating the parameters α and θ for gamma rounded data.

A subroutine in the main program calculated the MLE for α and θ using (7.2–6). In this subroutine, a Nag routine (DOIAJF) was used to compute an approximation to the integrals. Algorithms from Applied Statistics were used to find values of such functions as the incomplete gamma (Lau 1980) and digamma (Bernardo 1976). Convergence to the MLE had occurred if the absolute values of the difference between successive estimates of α and θ were both less than $E = 10^{-5}$.

To check the validity of the algorithm and subroutine, the following procedure was adopted. Using a Nag Quasi–Newton algorithm (EO4JAF) the MLE were found for various samples of rounded data. For these samples the MLE from the EM subroutine and Quasi–Newton agreed.

Simulation Results

Throughout the rest of this chapter the ML estimators for unrounded and rounded data will be denoted by $(\hat{\alpha},\hat{\theta})$ and $(\hat{\alpha}_R,\hat{\theta}_R)$ respectively. We define the efficiency of the ML estimators for rounded data as follows:

$$e(\hat{\alpha}_R,\hat{\alpha}) = \frac{MSE(\hat{\alpha})}{MSE(\hat{\alpha}_R)} \quad , \quad e(\hat{\theta}_R,\hat{\theta}) = \frac{MSE(\hat{\theta})}{MSE(\hat{\theta}_R)}$$

As in section (6.3) these efficiencies may be considered as a measure of the loss of information due to rounding.

Again as in section (6.4) the MLEs are conditional, the condition being existence. Where non-existence of MLE occurred, only samples where MLEs existed for both rounded and unrounded data were used to calculate the MSE.

The results from the simulation clearly showed the effect of the parameter α in influencing the efficiency of the MLE for rounded data. Generally as α decreased in value, so did the efficiency of the MLE. The loss in efficiency was greatest in $\hat{\alpha}_R$. The biases in $\hat{\alpha}_R$ and $\hat{\theta}_R$ were of the same order; it was the larger variance in $\hat{\alpha}_R$ that caused the loss in efficiency to be greater. Table (7.2.1) gives a selection of results for r = 1.0 and 1.5. Table (7.2.2) gives some indication of the possible influence of the lattice position on the efficiency of $\hat{\alpha}_R$ and $\hat{\theta}_R$. The effect of the lattice was seen to lessen as α increased in value. This was not unexpected, since the effect of rounding decreases as the distribution becomes less skewed.

Table 7.2.1

Efficiency of ML estimators $\hat{\alpha}_R$ and $\hat{\theta}_R$

		n = 25		n = 100		n = 500	
r	α	$e(\hat{\alpha}_R,\hat{\alpha})$	$e(\hat{\theta}_R,\hat{\theta})$	$e(\hat{\alpha}_R,\alpha)$	$e(\hat{\theta}_R,\theta)$	$e(\hat{\alpha}_R,\alpha)$	$e(\hat{\theta}_R,\hat{\theta})$
1.0	1	41.3	56.4	41.9	58.0	43.6	60.9
	5	66.4	69.1	70.9	81.9	73.9	80.9
	10	73.1	73.8	80.6	82.4	85.1	83.2
	20	76.4	77.8	84.1	84.5	87.6	85.6
1.5	1	26.1	43.2	27.7	44.9	29.8	45.7
	5	43.2	61.4	50.7	66.2	59.8	66.7
	10	58.7	63.9	59.2	65.4	60.3	67.9
	20	65.5	64.2	64.8	65.9	65.6	69.2

Note: all efficiencies in table at lattice position c = 0.

Table 7.2.2

Efficiency of ML estimators $\hat{\alpha}_R$ and $\hat{\theta}_R$ for n = 100, r = 1.0 and three lattice positions.

		c = 0	c = 0.25	c = 0.5
α = 1	$e(\hat{\alpha}_R,\hat{\alpha})$	41.9	28.4	17.4
	$e(\hat{\theta}_R,\hat{\theta})$	58.0	47.6	35.7
α = 5	$e(\hat{\alpha}_R,\hat{\alpha})$	70.9	71.4	67.8
	$e(\hat{\theta}_R,\hat{\theta})$	81.9	81.1	79.4

For the sample sizes and lattice positions considered, the results indicated that efficiencies of $\hat{\alpha}_R$ and $\hat{\theta}_R$ in the range 85–95% could be expected only for $\alpha \geq 5$

where $r \leq 0.5$. For $\alpha = 1$, the range was 65–80% when $r \leq 0.25$. This is in sharp contrast to the story for normal rounded data, where the efficiencies of MLEs of μ and σ^2 were in the range 85–92% for $r \leq 1.0$.

For coarse rounding of $r = 2.0$, the efficiency was as low as 6% and 33% for $\hat{\alpha}_R$ and $\hat{\theta}_R$ respectively, for sample sizes as large as 500, when $\alpha = 1$. At $\alpha = 20$ the efficiency for the same n and r could be as low as 49% and 53% for $\hat{\alpha}_R$ and $\hat{\theta}_R$ respectively.

We would expect the rate of convergence of the EM for the gamma to be slower than for the normal, the reason simply being that in a skewed distribution the loss of information caused by rounding is generally greater. As the rate of convergence of the EM is proportional to the loss in information, this will result in a slower convergence. Average numbers of iterations taken to converge when $E = 10^{-5}$ are given in Table (7.2.3) for the gamma and normal EM algorithms, when $n = 100$.

Table 7.2.3

Range of the average number of iterations for n = 100 across c = 0, 0.25 and 0.5.

	Average number of iterations		
r	Gamma		Normal
1	$\alpha < 5$	7 - 31	2.5
	$\alpha \geq 5$	6 - 7	
1.5	$\alpha < 5$	10 - 71	3.1
	$\alpha \geq 5$	8 - 10	

Table (7.2.3) shows that the gamma uses more iterations than the normal does. How much more expensive this may be is largely determined by the value of α.

7.2.2 Approximate Maximum Likelihood Estimates

Tallis (1967) gave methods for obtaining the approximate MLE for grouped data. The method was a slight but convenient modification of the results of Lindley (1950). The modification was to replace the various terms in Lindley's result by their expectations and obtain the average bias caused by rounding.

Using Tallis' method on the p.d.f. given by (7.2-1) we obtain the following approximate ML estimators for α and θ, for $c \neq 0$ and $\alpha > 2$.

$$\hat{\alpha} = \alpha_0 + \frac{\alpha_0 w^2}{24}\left[\theta_0^2(\alpha_0-1)(\alpha_0-2)[\alpha_0\psi'(\alpha_0)-1]\right]^{-1}$$
$$\hat{\theta} = \theta_0 - \frac{w^2}{24}\left[\theta_0(\alpha_0-1)(\alpha_0-2)[\alpha_0\psi'(\alpha_0)-1]\right]^{-1} \qquad (7.2\text{-}7)$$

α_0 and θ_0 are the usual MLEs obtained from the midpoints of the rounding intervals and $\psi'(z)$ is the trigamma function.

The restriction on c is obvious. If $c = 0$, then there is a possibility that the first rounding interval with midpoint zero has a non zero frequency. Then α_0 and θ_0 will not exist. The probability that α_0 and θ_0 do not exist is dependent on n, r and α.

Table 7.2.4

Probability that MLE α_0, θ_0 do not exist at c = 0.

α	r	n = 10	n = 100	n = 500
1	0.1	0.39	0.99	1.00
	0.25	0.72	1.00	1.00
5	1.0	0.06	0.43	0.94
	2.0	0.51	1.00	1.00
10	2.0	0.01	0.14	0.53
15	2.0	0.00	0.00	0.01

As shown by the results in Table (7.2.4) an increase in n or r, or a decrease in α, will cause the probability to rise. As n tends to infinity this probability will approach one. The results indicate that for $\alpha < 15$ where $r \leq 2.0$, there is a probability that α_0 and θ_0 will not exist, for samples as large as 500.

The restriction on α is caused by the fact that to obtain (7.2–7) we require the expected values of the reciprocals of x and x^2. These only both exist for $\alpha > 2$.

Usually the main advantage of the approximate ML method is that it is numerically simpler than the full ML method. However, for the gamma distribution, this should be put against the fact that the approximate ML method can only be used for $\alpha > 2$, and there is a definite probability at c = 0 that the estimates do not exist. Because of the disadvantages of this method of estimation it was decided not to investigate it any further.

7.2.3 Sheppard's Method

For the gamma distribution, Sheppard's method has obvious advantages over the ML approach, namely it is simple to use and it requires no iterations. From section (6.3) it has been shown that for $r \leq 1.5$ there is little difference between the efficiency of Sheppard's and ML methods for estimating the parameters of a normal distribution from rounded data. Even for $r \leq 2.0$ and $n \leq 100$ it may be of little preference to use the ML method. In this section we consider if such a region exists in the (r,n) plane for the gamma distribution.

$\underline{X} = (X_1,\ldots,X_n)$ is a random sample of size n from a gamma random variable X with parameters α and θ. Let $\underline{X}_R = (X_{R1},\ldots,X_{Rn})$ be the rounded sample where X_{Ri} is the rounded value of X_i corresponding to a rounding lattice with interval of width w and lattice position c. The estimates of α and θ by Sheppard's method are:

$$\tilde{\alpha} = \frac{\bar{X}_R^2}{S_R^2 - \frac{w^2}{12}} \qquad \tilde{\theta} = \frac{S_R^2 - \frac{w^2}{12}}{\bar{X}_R} \qquad (7.2\text{-}8)$$

where $\bar{X}_R$ and S^2_R are the usual estimators of the mean and variance applied to midpoints of the rounding intervals. $(\tilde{\alpha},\tilde{\theta})$ will be called Sheppard estimators of (α,θ).

In order to compare Sheppard and ML methods the efficiency of a Sheppard estimator will be defined as the MSE of the ML estimator divided by the MSE of the estimator in question. For unrounded data, we are simply comparing the

efficiency of the moment estimator with that of the ML estimator.

In the gamma distribution for unrounded data, the moment estimators are not appreciably less efficient than ML estimators in small to moderate size samples, though they are in large samples (Lawless 1982). The main purpose of the simulation study in this section is to investigate the situation for rounded data and to see if there is a region in the (r,n) plane where the moment estimators with Sheppard's corrections have a similar efficiency to the ML estimators.

Simulation Results

$\alpha = 1$

For this value of α the efficiency of Sheppard estimators could be very poor. At r = 1, the efficiency of these estimators could be as low as 27% for n = 100 and 2% at n = 500. Even with r = 0.25 the efficiency of $\tilde{\alpha}$ and $\tilde{\theta}$ could be as low as 60% and 63% respectively for n = 100. The larger MSE of Sheppard estimators was mainly caused by a higher bias in these estimators than in ML estimators. As expected an increase in r caused this bias to enlarge and consequently the efficiency to fall. The poor efficiency of Sheppard estimators at $\alpha = 1$ is not surprising. The results of Chapter 2 indicated that Sheppard Corrections can be unreliable in adjusting the moments of rounded data when the distribution is skewed.

Region $\alpha \geq 5$

Simulation results were obtained for α = 5, 10 and 20, and any discussion refers to this set of values.

In this region of α the efficiency of Sheppard estimators was greater than for α = 1. This is to be expected as Sheppard Corrections are generally more reliable as the distribution becomes less skewed. For $r \leq 1.0$ the bias in Sheppard estimators was of the same order as in the ML estimators. It was the greater variance in the Sheppard estimators that caused them to be less efficient. However for coarse rounding ($r > 1.0$) both the biases in $\tilde{\alpha}$ and $\tilde{\theta}$ and their variance were greater than for ML estimators. This resulted in a sharper decrease in efficiency, which was especially noticeable for $n > 100$.

Simulation results indicated:

(i) For $5 \leq \alpha \leq 20$ and $r \leq 1.0$, Sheppard estimators were at least 85% as efficient as ML estimators for sample sizes 500 or less.

(ii) For $5 \leq \alpha \leq 20$ and $r \leq 1.5$, Sheppard estimators were at least 75% as efficient as ML estimators for sample sizes 100 or less.

Although the study was limited for values of α upto 20, the results indicated that (i) and (ii) would hold for all values of α of 5 or more.

In regions of (r,n) in (i) and (ii), it may be preferable to use Sheppard's methods instead of the ML method. The loss in efficiency in using Sheppard estimators

may be worthwhile in view of their computational simplicity.

7.2.4 Naive Methods

In this section we briefly investigate the naive methods, ie method of moments and ML applied to the midpoints of the rounding intervals.

Define $\underline{X}_R = (X_{R_1},\ldots,X_{Rn})$ as in section (7.2.3). The naive ML estimators $(\hat{\alpha}_N,\hat{\theta}_N)$ are obtained by solving (7.2–9)

$$\log(\hat{\alpha}_N) - \psi(\hat{\alpha}_N) - \log(\bar{X}_R) + \log(\prod_i X_{Ri}) = 0$$
$$\hat{\theta}_N = \bar{X}_R/\hat{\alpha}_N \qquad (7.2\text{-}9)$$

The naive moment estimators $(\tilde{\alpha}_N,\tilde{\theta}_N)$ are given by

$$\tilde{\alpha}_N = \frac{\bar{X}_R^2}{S_R^2} \qquad \tilde{\theta}_N = \frac{S_R^2}{\bar{X}_R} \qquad (7.2\text{-}10)$$

Naive ML estimators

As mentioned in section (7.2.2) if c = 0 there is a probability that the MLE do not exist. This is a restriction on this method of estimation.

The results from the simulation showed the inconsistency of the ML approach applied to the midpoints. As n increased in size the expected values of the estimators became increasingly off target. Table (7.2.5) illustrates this point for

$\alpha = 5$.

Table 7.2.5

Expected values of Naive ML estimators and ML estimators, for $\alpha = 5$, r = 1.0 and c = 0.5.

	Naive ML estimators		ML estimators	
n	$E[\hat{\alpha}_N]$	$E[\hat{\theta}_N]$	$E[\hat{\alpha}_R]$	$E[\hat{\theta}_R]$
25	4.94	1.13	5.99	0.94
50	4.52	1.16	5.41	0.97
100	4.35	1.17	5.19	0.98
500	4.23	1.19	5.02	1.00

Because of the bias in $(\hat{\alpha}_N, \hat{\theta}_N)$, the MSE of these estimators could be considerably larger than ML estimators $(\hat{\alpha}_R, \hat{\theta}_R)$. Hence the efficiencies of the estimators $(\hat{\alpha}_N, \hat{\theta}_N)$ relative to $(\hat{\alpha}_R, \hat{\theta}_R)$ were generally poor unless the rounding was relatively low and sample size not very large. This was illustrated by the simulation results when they indicated that for $\alpha \geq 1$, this efficiency was 80% or more only when $r \leq 0.25$ and $n \leq 100$.

Naive Moment Estimators

Once again the simulation results demonstrated that a naive method of estimation can lead to misleading results. As n increased the expected values of $(\tilde{\alpha}_N, \tilde{\theta}_N)$ became increasingly off target. This was especially so for $\tilde{\alpha}_N$. The results showed that $E[\tilde{\alpha}_N] < E[\tilde{\alpha}]$ and $E[\tilde{\theta}_N] > E[\tilde{\theta}]$. Also in both cases the variances of (α_N, θ_N) were generally lower than $(\tilde{\alpha}, \tilde{\theta})$. Table (7.2.6) illustrates these points for

$\alpha = 5$. For certain values of n and r the naive estimators could be at least as efficient in terms of MSE's as the Sheppard estimators. For example, this was true for $\alpha \geq 1$ when $r \leq 1.0$ and $n \leq 50$.

Table 7.2.6

Mean and variance of Naive Moment and Sheppard estimators for $\alpha = 5$, r = 1.0 and c = 0.5.

	Naive Moment Estimators				Sheppard Estimators			
n	$E[\tilde{\alpha}_N]$	$E[\tilde{\theta}_N]$	$V[\tilde{\alpha}_N]$	$V[\tilde{\theta}_N]$	$E[\tilde{\alpha}]$	$E[\tilde{\theta}]$	$V[\tilde{\alpha}]$	$V[\tilde{\theta}]$
25	5.30	1.04	3.23	0.11	5.64	1.00	4.63	0.125
50	4.93	1.06	1.27	0.06	5.29	1.00	1.76	0.06
100	4.76	1.07	0.58	0.03	5.13	1.00	0.80	0.03
500	4.63	1.09	0.11	0.01	5.01	1.00	0.15	0.01

7.3 Exponential Distribution

This section briefly discusses the five methods of estimation of an exponential random variable X with unknown parameter θ, where X has been rounded. The p.d.f. for the exponential distribution is given by:

$$f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \qquad x \geq 0, \quad \theta > 0 \qquad (7.3\text{-}1)$$

Kulldorff (1961) devoted much of his book to the estimation of θ in the exponential distribution. He studied completely or partially grouped data with grouping having a finite number of intervals, an infinite number of intervals and

intervals of equal and unequal length. Where the data are rounded we are interested only in an infinite number of intervals, where the first group containing zero may be unequal from the other groups. For Kulldorff's estimators of θ for grouped data, the lattice position c was equal to a half. To limit the study of exponential rounded data, only that same value of c will be considered. Where necessary, reference will be made at other lattice positions.

7.3.1 Maximum Likelihood Estimation

From Kulldorff (1961) the ML estimator of θ for (7.3-1), where the sample data has been grouped into equal widths w and finite number k, is given by:

$$\hat{\theta}_R = w \Big/ \log\left[1 + \frac{n - n_k}{\sum_{i=2}^{k} (i-1)n_i}\right] \qquad (7.3\text{-}2)$$

with asymptotic variance $\frac{\theta^4}{nw^2}\left[1-e^{-\frac{w}{\theta}}\right]^2\left[e^{-\frac{w}{\theta}}-e^{-\frac{kw}{\theta}}\right]^{-1}$

where n_i is the number of observations in the ith group (i=1,...,k) and

$n = \sum_{i=1}^{k} n_i$.

For rounded data where k is infinite, (7.3-2) can be written as follows for a rounding interval w and lattice position $c = \frac{1}{2}$:

$$\hat{\theta}_R = \frac{w}{2 \tanh^{-1}\left[\frac{w}{2\bar{X}_R}\right]} \qquad (7.3\text{-}3)$$

The asymptotic variance is now $\frac{4\theta^4}{nw^2} \sinh^2\left[\frac{w}{2\theta}\right] = \frac{4\theta^2}{nr^2} \sinh^2\left[\frac{r}{2}\right]$

where $\bar{X}_R$ is the mean of the rounded sample $\underline{X} = (X_{R_1},\ldots,X_{Rn})$.

The expression (7.3–3) is the same as that given in Tricker (1984a). For $c \neq \frac{1}{2}$ an explicit form for $\hat{\theta}_R$ cannot be found and the likelihood equation has to be solved iteratively for $\hat{\theta}_R$.

Kulldorff (1961) investigated the properties of $\hat{\theta}$ for moderate n. It can be established from his results that the asymptotic results can be safely used for $n \geq 25$ when $r \leq 2.0$. As for the gamma distribution the efficiency of $\hat{\theta}_R$ can be defined as:

$$e(\hat{\theta}_R, \hat{\theta}) = \frac{MSE(\hat{\theta})}{MSE(\hat{\theta}_R)}, \text{ which is approximately equal to } \frac{r^2}{4 \sinh^2\left[\frac{r}{2}\right]} \qquad (7.3\text{-}4)$$

for $n \geq 25$, from Kulldorff's results. Table (7.3.1) shows the approximate efficiency for $\hat{\theta}_R$, compared with the efficiency of the parameters of normal and gamma distributions. The results in this table indicate that the ML estimators for gamma rounded data are considerably less efficient than those for normal and exponential rounded data. The results suggest that the efficiencies of the ML estimators of θ and μ are very similar.

Table 7.3.1

	Exponential $n \geq 25$	Normal all n		Gamma n = 10 $\alpha = 1$		$\alpha = 10$	
r	$e(\hat{\theta}_R, \hat{\theta})$	$e(\hat{\mu}_R, \hat{\mu})$	$e(\hat{\sigma}^2_R, \hat{\sigma}^2)$	$e(\hat{\alpha}_R, \hat{\alpha})$	$e(\hat{\theta}_R, \hat{\theta})$	$e(\hat{\alpha}_R, \alpha)$	$e(\hat{\theta}_R, \hat{\theta})$
1	92.1	92.3	85.5	17.4	35.7	77.6	78.4
1.5	83.2	84.2	71.8	12.3	25.6	64.1	68.9
2.0	72.4	74.6	54.1	17.3	17.3	62.4	62.6

Note: all efficiencies in table at lattice position c = $\frac{1}{2}$.

7.3.2 Other Methods of Estimation

Tallis (1967) gave the approximate ML estimator of θ for exponential rounded data. At c = $\frac{1}{2}$, the ML estimator (7.3-3) is simple to obtain, thus the approximate ML estimator has no advantage. However for c $\neq \frac{1}{2}$ this is not so and the approximate ML method may be a suitable approximation to the exact MLE until rounding becomes quite coarse.

For exponential rounded data the naive method of moments and ML, together with Sheppard's method, all produce the same estimator of θ, namely

$$\tilde{\theta} = \bar{X}_R \qquad (7.3\text{-}5)$$

This will be referred to simply as the Sheppard estimator of θ.

Using the expressions for $E[\bar{X}_R]$ and $V[\bar{X}_R]$ from Tricker (1984a), the MSE of $\tilde{\theta}$ at $c = \frac{1}{2}$ is given by:

$$MSE[\tilde{\theta}] = \frac{\theta^2}{n}\left[\frac{r^2e^{-r}}{(1-e^{-r})^2}\right] + \theta^2\left[1 - \frac{r}{2} + \frac{re^{-r/2}}{e^{r/2}-e^{-r/2}}\right]^2 \qquad (7.3\text{-}6)$$

For $n \geq 25$, the approximate efficiency of the Sheppard estimator $\tilde{\theta}$ relative to ML estimator $\hat{\theta}_R$ is the asymptotic variance of $\hat{\theta}_R$ divided by the MSE($\tilde{\theta}$). Some of these efficiencies are given in Table (7.3.2). The results in this table suggest that, for $r \leq 0.5$ and $n \leq 100$, Sheppard's and ML methods are equally efficient. Outside this region of (r,n) Sheppard's method is far less efficient. For $c \neq \frac{1}{2}$ we would expect efficiencies similar to those given in Table (7.3.2).

Table 7.3.2

Efficiency of $\tilde{\theta}$ relative to $\hat{\theta}_R$ for $c = \frac{1}{2}$

	$e(\tilde{\theta}, \hat{\theta}_R)$		
r	n = 25	n = 100	n = 500
0.5	100	99.7	85.4
1.0	99.8	68.2	25.5
1.5	4.0	1.0	0.2

7.4 Discussion of Results

In this chapter estimation procedures for rounded data from non-normal distributions have been considered. For the gamma distribution the EM algorithm was used to obtain the MLE of the parameters, for rounded data. The loss in

efficiency of the MLE due to rounding was larger than for the normal distribution. This was especially so for $\alpha < 5$, where the rounding was coarse ($r > 1$). For $\alpha \leq 5$, we could expect losses in efficiency between 5–15% for rounding with $r \leq 0.5$. In contrast, for normal rounded data such losses in efficiency would not take place until $r = 1.0$.

In Chapter 6, Sheppard's method was found to be competitive with the ML approach in terms of efficiency, for normal rounded data. This was found to be less so for gamma rounded data, especially when the distribution is very skewed. However, as shown in section (7.2.3) there are regions in the (r,n) plane where it may be preferable to use Sheppard's method instead of the ML method. The loss in efficiency in using Sheppard estimators being counter balanced by their simplicity.

The approximate MLE could be found in only a restricted range of α and may not exist when $c = 0$. As a result this method of estimation was limited for rounded gamma data.

The naive methods of estimation for gamma rounded data had limited use. The naive MLE had a problem of existence at $c = 0$, and was found to have poor efficiency relative to MLE unless the rounding precision r was low and sample size not very large. The naive moment estimators were found to be at least as efficient as Sheppard estimators for only $r \leq 1.0$ and $n \leq 50$. Hence Sheppard estimators are preferable as they have higher efficiency over a large range of r and n.

For the exponential distribution the loss in efficiency of the MLE due to rounding was considerably less than for the gamma. In fact the efficiency of the ML estimators of θ and μ in the exponential and normal respectively were almost identical. As in the gamma the efficiency of Sheppard estimators relative to ML estimators were poor outside a limited region of r and n.

Several points need to be made in the light of the results in this chapter.

1. For the normal distribution the loss in efficiency in MLE due to rounding is small. This gives an optimistic impression of the effect of rounding on the method of maximum likelihood. The results in this chapter indicate that for non-normal distributions the loss in efficiency in the MLE when the data has been rounded can be considerable.

2. For normal rounded data simpler methods of estimation than ML were as efficient for very coarse rounding. The results in this chapter suggest that, for other distributions, the ML method will be generally more efficient than other standard methods of estimation.

3. As shown in Chapter 6, when the EM algorithm is applied to normal rounded data it is not very expensive in terms of number of iterations for $r \leq 2.0$. The results in this chapter indicate that this is not necessarily true for other distributions.

The aim of this study was to examine the effect of rounding on certain statistical procedures. The question has been investigated whether or not one should be concerned about the degree of precision of the recorded data. The study has illustrated the suitability of certain statistical methods to rounded data.

In the literature there is a large amount of scattered information concerned with rounding. It became evident that a coherent survey of this work was needed. Chapter 1 of this thesis contains the first major comprehensive literature review on the topic of rounding. It soon became obvious that certain areas had been thoroughly investigated while others, notably the consequences of rounding on tests of significance and suitability of estimation procedures for rounded data, had been neglected.

Most of the early statistical literature on rounding dealt with the derivation of relationships between the moments calculated from data before and after rounding. Chapter 2 is concerned with this relationship. Adapting the quantization theory from communication engineering, the characteristic function of the rounded distribution X_R is obtained. A proof of the characteristic function of X_R is presented which is much simpler and more elegant than the one presented in quantization theory. For univariate distributions via the characteristic function of X_R, explicit expressions for the moments of X_R are obtained for the first time. These are used to determine the bias in the moments of a distribution caused by rounding. This has been the most extensive study to date, as it has considered not only the degree of rounding, but also lattice position and shape of the distribution. The results show how departure from symmetry was a crucial factor in deciding the size of the rounding bias in the moments. Generally as the distribution becomes increasingly non-symmetrical, rounding bias increases. The

degree of precision r which render this bias negligible is dependent on the shape of the distribution. In the final part of the chapter it is shown how the theoretical results can be extended to higher dimensional rounded random variables. A new result giving an exact expression for the joint first moment of a bivariate distribution is derived. It is shown how this expression can be further simplified if the characteristic function exhibits certain symmetric properties. Attention was focussed on the joint first moment of the bivariate normal distribution. The rounding bias in this moment was found to be dependent on the correlation. As the correlation increased so did the bias.

Over the years Sheppard's corrections have been universally regarded as the acceptable method for determining the relationship between the moments of X and X_R. For the normal distribution these corrections were found to be a reasonable approximation to the moments of rounded data for $r \leq 2.0$. However departure from normality can result in these corrections becoming very unreliable. Further, as illustrated by the bivariate normal, Sheppard's corrections can be unreliable in adjusting the joint first moment for rounded data. It is concluded that these corrections should be handled with care.

Whereas the previous literature is informative on the behaviour of the moments for rounded data, the effect of rounding on tests of significance has been generally unexplored. There is considerable vagueness in the literature concerning what level of precision should be used when applying a statistical test. In Chapters 3 and 4 the effect of rounded normal data on the significance level and power of five normal test statistics is investigated.

Generally it is impossible to obtain an explicit form for the sampling distribution of the test statistic for rounded data. The approach under this constraint was to combine extensive simulation with approximations to the sampling moments of the test statistics, in order to examine the effect of rounding on a test statistic. This approach was found to be very successful in determining the performance of a test statistic with respect to the significance level and power for rounded data. The results showed that tests of means are more robust under rounding than tests of variances. The two sample t–test and F–test in the anlysis of variance are found to be the least affected by rounded data, while the chi–squared test for a variance is the most sensitive. For tests of means the power under rounding can be approximated using a non–central distribution with a given adjustment to the non–central parameter. This provides a good estimate of the power under rounding. It is also shown in Chapter 4 how the chi–squared test can be made more robust to rounding by making a simple adjustment to the test statistic. Tests of hypothesis regarding the value of variance should be based on this adjusted test statistic if the data is subject to rounding.

In the literature various rules have been suggested for the degree of precision that should be used when recording data. There seems to be no standard rule. There is a need to know when normal theory tests can be applied 'safely' to rounded data. In Chapters 3 and 4 guidance is given on what is an appropriate degree of precision to use when applying certain tests to normal rounded data. The results show that we can use far less precision in the data than originally realised, without the significance level and power of the test being adversely affected. The usual rules of rounding presented in the literature were generally found to be too conservative.

In Chapter 2 it is shown how non-normality can increase the effect of rounding. In many situations the tests considered in Chapters 3 and 4 are used where the assumption of normality is invalid. It is only sensible to investigate how the tests will perform for rounded non-normal data. The results of such an investigation are given in Chapter 5. Guidelines are given on how the previous degree of precision recommended for normal populations can be applied when the population is non-normal. A notable result of this investigation is the robustness of the two sample t-test and F-test in the analysis of variance over a large number of rounded non-normal populations. The results of Chapter 5 give, for the first time, an indication of how much non-normality can be allowed before the effect of rounding distorts the significance level and power of certain normal tests.

The effect of rounding on both the moments and test statistics can be increasd by the departure from normality of the population. The last two chapters of this thesis are concerned with how standard estimation procedures perform in terms of efficiency for normal and non-normal rounded data.

For normal rounded data simpler methods of estimation than maximum likelihood are found to be as efficient for coarse rounding. Moment estimators with Sheppard's corrections are found to be competitive in terms of efficiency as maximum likelihood estimators. However, evidence from the non-normal distributions considered suggests that, if rounding is coarse or the distribution is very skewed, the maximum likelihood approach is preferable in terms of efficiency.

Previous research has shown that the loss in efficiency in the maximum likelihood estimate due to rounding is small. That research was restricted to the normal distribution. This has given an optimistic impression of the effect of rounding on

the method of maximum likelihood. Our results indicate that the loss in efficiency can be considerable for non–normal rounded data.

To find the maximum likelihood estimates from normal and gamma rounded data the EM algorithm was used. This is the first time this algorithm has been applied to non–normal rounded data. For normal rounded data the EM algorithm was not very expensive in terms of the number of iterations required. However for rounded gamma data far more iterations were needed. The results indicate that when the rounding is coarse or the distribution very skewed the rate of convergence of the EM algorithm can be slow. This is an important point to consider when deciding on the suitability of this algorithm for rounded non–normal data.

Further Research

Throughout the course of this study, it has become apparent that there is further research to be carried out into the subject of rounded data. In Chapter 5 only sample sizes of n = 25 are considered. There is evidence that when the departure from normality is extreme, increasing the sample size can cause rounding to have a greater effect on the significance level and power of a test. Further work in this area is worthy of investigation.

In this thesis it has been shown that normal test statistics are more robust to rounded data than previously realised. However, what about other statistical tests? The difficulty of applying tests based on ordered data, such as Shapiro–Wilk test and Spearmans rank correlation test, is that sample data may contain ties resulting from rounding. This problem of ties can disturb the sampling distribution of the

test statistic. Further work on what is an appropriate degree of precision for the recorded data for such tests is required.

It is clear from the work in this thesis that departures from normality in a population generally increase the effect of rounding. The impact of rounding on non-normal regression models needs further investigation. The problems in estimating the parameters of multivariate distributions when the data has been subject to rounding, have not been well covered. However a possible difficulty in this area is that the maximum likelihood method may be restricted, this being caused by the difficulties encountered in evaluating the necessary integrals.

Model selection is another area which needs further investigation when dealing with rounded data. Techniques such as plotting procedures and tests for examining a model's suitability for rounded data could be included in this investigation.

One of the major difficulties when dealing with rounded data is that there is a lack of relevant statistical theory which can be applied. As a result we often have to look at specific cases. Another problem is that the application of statistical methods on rounded data often requires numerical techniques. In the past, progress on analysing rounded data has been restricted by the computational effort associated with these techniques. However, with advances in computing, this is now becoming less of a problem. Thus we would expect further development in rounded data research.

APPENDIX A : COMPUTER PROGRAMS AND OUTPUT FOR CHAPTER 2

Several FORTRAN programs were written to investigate the effect of rounding on the moments of a probability distribution. All the programs were run under the FORTRAN 77 compiler on the IBM Mainframe computer at Sheffield City Polytechnic.

The main program in this chapter was JMOMENTS. This program obtains by Monte Carlo methods an estimate of the mean, variance, skewness and kurtosis of data from a Johnson distribution which has been subject to rounding. The following is a list of all the output produced by this program, upon which the contour diagrams Figures (2.2.10) to (2.2.13) are based.

The four parameters μ_R, σ^2_R, $\sqrt{\beta_1}_R$ and β_{2R} were estimated for a series of Johnson distributions whose shape parameters $\sqrt{\beta_1}$ and β_2 fell on a grid with $\sqrt{\beta_1}$ = 0.0, 0.2, 0.4, 0.6, 0.8 and β_2 = 1.7, 2.0, 2.4, 2.8, 3.2, 3.6, 4.4. Omitting β_2 = 1.7 where $\sqrt{\beta_1}$ = 0.6 and 0.8, produced 33 distributions. For each combination of $(\sqrt{\beta_1},\beta_2)$, the four parameters were found for r = 1.5, 1.0 and 0.5, each of 11 values of c ie −0.5, −0.4, ..., 0.4, 0.5.

A listing of all the output mentioned above is available on request.

APPENDIX B : COMPUTER PROGRAMS AND OUTPUT FOR CHAPTERS 3 AND 4

In this appendix computer programs written to find the significance level and power of a test for rounded normal data are considered. A list of output produced by these programs is provided. This appendix also contains tables of results for α_R (significance level of test under rounding) which are referred to in Chapter 3.

Several FORTRAN programs were written to investigate the effect of rounded normal data on the significance level and power of a test. All the programs were run under the FORTRAN 77 compiler on the IBM Mainframe computer at Sheffield City Polytechnic. A list of the purpose written programs used are given below.

(i) EXACT – this program calculates the exact value of the significance level of 4 standard tests for rounded normal data for various combinations of r, n and α, over a range of c values. The tests were the one and two sample t–tests, chi–squared test and F–test.

(ii) SIMUL – this program obtains an estimate of the significance level of 6 standard tests for rounded normal data by Monte Carlo Methods. The significance level can be obtained for various combinations of r, n and α, over a range of c values. The tests were: one and two sample t–tests, chi–squared test and F–test, together with the F–test in the one and two way analysis of variance.

(iii) PEXACT – this program calculates the exact value of the power of the 4 standard tests given in (i), for rounded normal data, for various combinations of r, n and α over a range of c values.

(iv) PSIMUL – this program obtains an estimate of the power of the 6 standard tests given in (ii) for rounded normal data by Monte Carlo Methods. This estimate of power can be obtained for various combinations of r, n and α, over a range of c values.

A listing of the above programs is available on request.

The following is a list of all the output produced by the four programs given in (i) – (iv), upon which the study of the effect of rounded normal data on the significance level and power of a test was based.

SIGNIFICANCE LEVEL OF A TEST

The significance level of each test under rounding was evaluated for values corresponding to the lower and upper 0.1%, 1.0%, 2.5% and 5% points under normal theory conditions with no rounding. For each combination of n, k and r given below the eight percentage points were found for 11 values of c, ie c = −0.5, −0.4, ..., 0.4, 0.5. The results from the SIMUL program were based on 100,000 iterations.

One sample t-test

EXACT: n = 2 to 15 with r = 0.25, 0.5, 1.5, 2.0

SIMUL: n = 25, 30 with r = 0.5, 1.0, 1.5, 2.0

Chi-squared test

EXACT: n = 2 to 15 with r = 0.25, 0.5, 1.0, 1.5, 2.0

SIMUL: n = 25, 26, 30 with r = 0.25, 0.5, 1.0, 1.5, 2.0

n = 50, 100 with r = 0.5

Two sample t-test

EXACT: n = 2 to 7 with r = 0.25, 0.5, 1.0, 1.5, 2.0

SIMUL: n = 10, 25 with r = 0.25, 0.5, 1.0, 1.5, 2.0

F-test

EXACT: n = 2 to 5 with r = 0.25, 0.5, 1.0, 1.5, 2.0

SIMUL: n = 10, 25 with r = 0.25, 0.5, 1.0, 1.5, 2.0

n = 40, 100 with r = 1.5, 2.0

One Way Analysis of Variance

SIMUL: k = 3, n = 5 to 10, 25 with r = 0.25, 0.5, 1.5, 2.0

k = 4, n = 5, 6, 10 with r = 0.25, 0.5, 1.5, 2.0

k = 5, n = 5, 10, 25 with r = 0.5, 1.5, 2.0

k = 10, n = 5, 10 with r = 0.5, 1.5, 2.0

Two Way Analysis of Variance

SIMUL: k = 3, n = 5, 10, 25 with r = 0.25, 0.5, 1.5, 2.0

k = 5, n = 5, 10 with r = 1.0, 1.5, 2.0

k = 10, n = 5, 10 with r = 1.0, 1.5, 2.0

POWER OF A TEST

The PEXACT program was only used to check the results of the PSIMUL program. All power results given in Chapter 4 were obtained by simulation using the PSIMUL program. The power of each test under rounding was evaluated for values of the alternative hypothesis H_1: (one tailed) corresponding to powers of 0.3, 0.5, 0.7 and 0.95 under normal theory conditions with level of significance α. For each combination of n, k, r and α given below, the four power levels were found for 11 values of c using the PSIMUL program. The estimate of each power was based on 100,000 iterations.

One sample t-test

$\alpha = 0.05$: n = 5 with r = 0.5, 1.5, 1.0

n = 10 with r = 1.0, 1.5

n = 25 with r = 1.0, 1.5

$\alpha = 0.01$ and 0.001: n = 10 with r = 1.0

n = 25 with r = 1.5

Chi-squared test

$\alpha = 0.05$: n = 5, 10, 25 with r = 0.25, 0.5, 1.0

Two sample t-test

$\alpha = 0.05$: n = 5, 10, 25 with r = 0.5, 1.0, 1.5, 2.0

$\alpha = 0.01$ and 0.001: n = 10, 25 with r = 1.5, 2.0

F-test

$\alpha = 0.05$: n = 5 with r = 0.25, 0.5, 1.0

n = 10 with r = 0.25, 0.5, 1.0, 1.5

One Way Analysis of Variance

$\alpha = 0.05$: $k = 3, n = 5$ with $r = 1.0, 1.5$

$k = 3, n = 10, 25$ with $r = 1.5, 2.0$

$k = 5, n = 5, 25$ with $r = 1.5, 2.0$

Two Way Analysis of Variance

$\alpha = 0.05$: $k = 3, n = 5$ with $r = 1.5$

$k = 5, n = 5$ with $r = 2.0$

A listing of all the output mentioned above is available on request.

TABLE OF RESULTS FOR α_R (Chapter 3)

Table B.1

Values of α_R (%) for one sample t-test for n = 5 when r = 2.0, 1.0 and 0.5

		α(%) lower tail				α(%) upper tail			
r	c	0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
2.0	-0.5	2.48	2.48	3.12	3.13	3.13	3.12	2.48	2.48
	-0.4	4.74	4.75	6.52	6.52	1.37	1.32	1.32	1.12
	-0.3	7.82	7.82	8.53	12.09	4.72	0.49	0.49	0.43
	-0.2	11.16	11.18	11.39	13.13	1.98	1.86	0.16	0.14
	-0.1	13.82	13.87	14.55	14.57	0.72	0.72	0.04	0.04
	0.0	0.01	0.22	0.23	2.09	2.09	0.23	0.22	0.01
	0.1	0.04	0.04	0.72	0.72	14.57	14.55	13.87	13.82
	0.2	0.14	0.16	1.86	1.98	13.13	11.39	11.18	11.16
	0.3	0.43	0.49	0.49	4.72	12.09	8.53	7.82	7.82
	0.4	1.12	1.32	1.32	1.37	6.52	6.52	4.75	4.74
	0.5	2.48	2.48	3.12	3.13	3.13	3.12	2.48	2.48
1.0	-0.5	0.47	0.87	2.70	4.13	4.13	2.70	0.87	0.47
	-0.4	0.59	1.26	3.89	4.52	4.24	2.01	1.72	0.36
	-0.3	0.70	1.07	3.86	6.44	6.77	1.29	1.17	0.29
	-0.2	0.78	1.37	2.57	5.62	4.63	2.21	0.73	0.20
	-0.1	0.86	1.42	3.54	4.96	3.25	2.96	0.47	0.13
	0.0	0.08	0.92	1.89	6.08	6.08	1.89	0.92	0.08
	0.1	0.13	0.47	2.96	3.25	4.96	3.54	1.42	0.86
	0.2	0.20	0.73	2.21	4.63	5.62	2.57	1.37	0.78
	0.3	0.29	1.17	1.29	6.77	6.44	3.86	1.07	0.70
	0.4	0.36	1.72	2.01	4.24	4.52	3.89	1.26	0.59
	0.5	0.47	0.87	2.70	4.13	4.13	2.70	0.87	0.47

Table B.1 (continued)

		α(%) lower tail				α(%) upper tail			
r	c	0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
0.5	-0.5	0.14	0.88	2.40	5.12	5.12	2.40	0.88	0.14
	-0.4	0.16	1.09	2.46	4.76	5.05	2.50	1.03	0.11
	-0.3	0.14	0.99	2.83	5.09	5.57	2.17	1.19	0.12
	-0.2	0.11	1.01	2.33	5.04	4.97	2.16	0.96	0.10
	-0.1	0.08	0.99	2.61	5.10	4.80	2.85	0.91	0.09
	0.0	0.07	0.90	2.43	5.18	5.18	2.43	0.90	0.07
	0.1	0.09	0.91	2.85	4.80	5.10	2.61	0.99	0.08
	0.2	0.10	0.96	2.16	4.97	5.04	2.33	1.01	0.11
	0.3	0.12	1.19	2.17	5.57	5.09	2.83	0.99	0.14
	0.4	0.11	1.03	2.50	5.05	4.76	2.46	1.09	0.16
	0.5	0.14	0.88	2.40	5.12	5.12	2.40	0.88	0.14

Table B.2

Minimum and maximum values of α_R(%) found for the 11 values of c, in a one sample t-test for n = 10

		α(%) lower tail				α(%) upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
2.0	min	0.01	0.37	1.22	2.03	2.03	1.22	0.37	0.01
	max	1.91	4.04	4.23	10.00	10.00	4.23	4.04	1.91
1.0	min	0.07	0.86	2.21	4.82	4.82	2.21	0.86	0.07
	max	0.15	0.12	2.71	5.60	5.60	2.71	0.12	0.15
0.5	min	0.09	0.96	2.44	4.95	4.95	2.44	0.96	0.09
	max	0.11	1.03	2.56	5.17	5.17	2.56	1.03	0.11

Table B.3

Minimum and maximum values of $\alpha_R(\%)$ found for the 11 values of c, in a one sample t–test for n = 25

		α(%) lower tail				α(%) upper tail				mean		variance	
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1	t	t_R	t	t_R
2.0	min	0.03	0.65	1.66	4.31	0.03	0.65	1.66	4.31	0	-0.01	1.09	1.09
	max	0.27	1.34	3.17	6.68	0.27	1.34	3.17	6.68	0	0	1.09	1.11
1.5	min	0.06	0.90	2.40	4.79	4.79	2.40	0.90	0.06	0	-0.01	1.09	1.09
	max	0.12	1.15	2.85	5.70	5.70	2.85	1.15	0.12	0	0	1.09	1.11

Table B.4

Minimum and maximum values of $\alpha_R(\%)$ found for the 11 values of c, in the chi–squared test for a varaince, for n = 5

		α(%) lower tail				α(%) upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
2.0	min	4.95	4.95	4.95	4.95	14.74	14.74	1.43	0.86
	max	14.85	14.85	14.85	14.85	15.43	15.43	2.12	1.11
1.0	min	0.94	0.94	0.94	0.94	5.98	3.54	1.32	0.20
	max	0.99	0.99	0.99	0.99	6.04	3.56	1.35	0.21
0.5	min	0.07	0.68	1.86	5.81	5.50	2.87	1.53	0.12
	max	0.07	0.68	1.86	5.81	5.51	2.88	1.60	0.12
0.25	min	0.13	0.77	2.47	5.34	5.10	2.53	1.02	0.09
	max	0.13	0.77	2.47	5.34	5.10	2.53	1.02	0.09

Table B.5

Minimum and maximum values of $\alpha_R(\%)$ found for the 11 values of c in the chi-squared test for a variance, for n = 10

		$\alpha(\%)$ lower tail				$\alpha(\%)$ upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
1.5	min	0.05	0.58	0.58	0.58	11.58	4.76	2.52	0.40
	max	0.24	2.11	2.11	2.11	12.84	6.04	2.77	0.43
1.0	min	0.06	0.44	2.05	3.32	6.72	3.74	1.75	0.23
	max	0.09	0.60	2.47	3.33	6.75	3.77	1.77	0.23
0.5	min	0.11	0.90	2.36	4.63	5.56	2.88	1.12	0.12
	max	0.11	0.90	2.36	4.63	5.56	2.88	1.13	0.12
0.25	min	0.11	0.91	2.45	4.89	5.17	2.56	1.02	0.11
	max	0.11	0.91	2.45	4.89	5.17	2.56	1.02	0.11

Table B.6

Minimum and maximum values of $\alpha_R(\%)$ found for the 11 values of c in the chi-squared test for a variance, for n = 25

		$\alpha(\%)$ lower tail				$\alpha(\%)$ upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
1.5	min	0.00	0.06	0.33	1.31	15.24	9.71	4.81	0.85
	max	0.06	0.37	0.95	2.16	15.46	10.72	5.21	0.94
1.0	min	0.05	0.51	1.43	2.88	8.74	5.04	2.23	0.30
	max	0.07	0.63	1.72	3.17	9.39	5.34	2.35	0.32
0.5	min	0.08	0.88	2.15	4.41	5.94	3.02	1.24	0.13
	max	0.11	0.90	2.23	4.49	6.00	3.10	1.29	0.14
0.25	min	0.09	0.95	2.45	4.87	5.21	2.63	1.03	0.11
	max	0.11	1.00	2.49	4.93	5.32	2.67	1.08	0.12

Table B.7

Minimum and maximum values of $\alpha_R(\%)$ found for 11 values of c in the two sample t–test, for n = 5

		α(%) lower tail				α(%) upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
1.0	min	0.07	0.89	2.53	4.81	4.79	2.50	0.88	0.07
	max	0.10	0.94	2.59	4.93	4.91	2.58	0.95	0.10
1.5	min	0.04	0.55	2.60	5.10	5.09	2.57	0.54	0.04
	max	0.10	0.95	2.92	5.32	5.33	2.89	0.93	0.09

Table B.8

Minimum and maximum values of $\alpha_R(\%)$ found for 11 values of c in the two sample t–test for n = 10 and 25 where r = 2.0

		α(%) lower tail				α(%) upper tail			
n		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
10	min	0.04	0.86	2.27	4.92	4.90	2.26	0.86	0.04
	max	0.14	1.02	2.45	5.26	5.24	2.44	1.06	0.14
25	min	0.09	0.98	2.51	4.90	4.89	2.49	0.96	0.09
	max	0.12	1.06	2.62	5.03	5.05	2.60	1.05	1.13

Table B.9

Minimum and maximum values of $\alpha_R(\%)$ found for 11 values of c in the F-test for both samples of size 5

		α(%) lower tail				α(%) upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
2.0	min	4.71	4.71	4.71	4.90	4.90	4.71	4.71	4.71
	max	12.64	12.64	12.64	12.78	12.78	12.64	12.64	12.64
1.0	min	0.93	1.10	2.05	4.94	4.94	2.05	1.10	0.93
	max	0.98	1.16	2.11	4.97	4.97	2.11	1.16	0.98
0.5	min	0.09	1.06	2.52	5.13	5.13	2.52	1.06	0.09
	max	0.10	1.10	2.56	5.23	5.23	2.56	1.10	0.10

Table B.10

Minimum and maximum values of $\alpha_R(\%)$ found for 11 values of c in the F-test for both samples of size 10

		α(%) lower tail				α(%) upper tail			
r		1.0	1.0	2.5	5.0	5.0	2.5	1.0	0.1
1.0	min	0.08	1.02	2.41	5.01	5.00	1.01	2.39	0.08
	max	0.11	1.08	2.61	5.22	5.20	1.08	2.60	0.11
0.5	min	0.10	1.03	2.51	5.01	5.01	2.50	1.01	0.10
	max	0.12	1.07	2.60	5.11	5.10	2.59	1.06	0.12

Table B.11

Minimum and maximum values of $\alpha_R(\%)$ found for 11 values of c in the F-test for both samples of size 25

		$\alpha(\%)$ lower tail				$\alpha(\%)$ upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
1.5	min	0.04	0.80	2.24	4.66	4.64	2.24	0.78	0.04
	max	0.15	1.08	2.58	5.06	5.05	2.56	1.07	0.14
2.0	min	0.09	0.97	2.41	4.80	4.83	2.42	0.10	0.09
	max	0.11	1.04	2.52	5.00	5.10	2.53	1.05	0.11

Tabler B.12

Minimum and maximum values of $\alpha_R(\%)$ found for the 11 values of c, in a One-way Analysis of Variance for k = 3 and n = 5, 10, 25

k = 3 n = 5

		$\alpha(\%)$ lower tail				$\alpha(\%)$ upper tail				mean		variance	
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1	F	F_R	F	F_R
2.0	min	5.20	5.20	5.20	5.20	3.97	2.21	0.45	0.05	1.20	1.19	2.16	2.05
	max	5.81	5.81	5.81	5.81	5.31	3.41	0.90	0.23	1.20	1.21	2.16	2.35
1.5	min	3.42	3.42	3.42	3.49	4.67	2.51	0.87	0.08	1.20	1.20	2.16	2.06
	max	3.56	3.56	3.56	3.63	4.85	2.79	0.93	0.16	1.20	1.21	2.16	2.33
1.0	min	1.66	1.63	1.69	4.39	4.66	2.45	1.00	0.10	1.20	1.20	2.16	2.13
	max	1.77	1.77	1.80	4.57	4.82	2.54	1.04	0.12	1.20	1.21	2.16	2.21
0.5	min	0.43	0.62	2.65	4.99	4.91	2.48	0.97	0.10	1.20	1.20	2.16	2.14
	max	0.49	0.69	2.79	5.08	5.06	2.56	1.06	0.12	1.20	1.21	2.16	2.21
0.25	min	0.10	0.97	2.49	5.03	4.91	2.49	1.02	0.11	1.20	1.20	2.16	2.16
	max	0.14	1.06	2.61	5.11	5.00	2.54	1.05	0.12	1.20	1.20	2.16	2.18

Table B.12 (continued)

k = 3 n = 10

		α(%) lower tail				α(%) upper tail				mean		variance	
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1	F	F_R	F	F_R
2.0	min	2.42	2.40	2.40	2.40	4.44	1.93	0.74	0.03	1.08	1.05	1.37	1.21
	max	2.81	2.81	2.82	2.84	5.09	2.70	1.04	0.14	1.08	1.08	1.37	1.41
1.5	min	1.56	1.56	1.63	3.05	4.46	2.20	0.82	0.05	1.08	1.05	1.37	1.25
	max	1.97	1.97	1.97	3.73	5.06	2.69	1.03	0.14	1.08	1.08	1.37	1.39
1.0	min	0.66	0.66	1.50	4.93	4.66	2.37	0.92	0.04	1.08	1.06	1.37	1.27
	max	1.03	1.03	1.92	5.42	5.06	2.55	1.08	0.13	1.08	1.07	1.37	1.33
0.5	min	0.14	0.93	2.14	4.46	4.63	2.26	0.91	0.05	1.08	1.06	1.37	1.27
	max	0.28	1.24	2.65	5.12	4.91	2.55	1.00	0.10	1.08	1.07	1.37	1.32
0.25	min	0.03	0.81	2.20	4.55	4.75	2.31	0.94	0.05	1.08	1.06	1.37	1.29
	max	0.12	0.95	2.45	4.89	4.83	2.52	1.01	0.09	1.08	1.06	1.37	1.30

k = 3 n = 25

		α(%) lower tail				α(%) upper tail				mean		variance	
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1	F	F_R	F	F_R
2.0	min	1.06	1.06	1.07	6.73	4.80	2.43	0.97	0.09	1.03	1.03	1.12	1.10
	max	1.15	1.15	1.16	7.12	5.04	2.57	1.01	1.11	1.03	1.03	1.12	1.13
1.5	min	0.68	0.68	2.54	4.71	4.85	2.45	0.95	0.10	1.03	1.03	1.12	1.11
	max	0.73	0.73	2.62	4.83	5.02	2.54	1.02	0.11	1.03	1.03	1.12	1.12
1.0	min	0.32	0.49	2.28	5.27	4.95	2.48	0.97	0.09	1.03	1.03	1.12	1.12
	max	0.36	0.55	2.40	5.43	5.07	2.56	1.02	0.12	1.03	1.03	1.12	1.13
0.5	min	0.08	0.90	2.42	4.92	4.96	2.47	0.95	0.09	1.03	1.03	1.12	1.11
	max	0.11	0.99	2.56	5.01	5.12	2.54	1.00	0.11	1.03	1.03	1.12	1.12
0.25	min	0.13	0.98	2.44	4.94	5.01	2.49	0.97	0.10	1.03	1.03	1.12	1.12
	max	0.16	1.07	1.07	5.06	5.05	2.54	1.01	0.11	1.03	1.03	1.12	1.12

Table B.13

Minimum and maximum values of $\alpha_R(\%)$ found for the 11 values of c, in a One-way Analysis of Variance for k = 5, 10 where n = 5

k = 5, n = 5

		α(%) lower tail				α(%) upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
2.0	min	0.37	0.43	1.80	5.26	4.44	2.12	0.73	0.09
	max	0.45	0.52	2.19	5.86	5.15	2.90	1.12	0.16
1.5	min	0.16	0.70	2.75	4.71	4.94	2.49	0.94	0.08
	max	0.17	0.88	2.84	4.88	5.24	2.60	1.04	0.12
1.0	min	0.03	0.98	2.39	4.85	5.03	2.56	1.00	0.09
	max	0.06	1.03	2.53	5.06	5.14	2.61	1.04	0.10
0.5	min	0.09	0.98	2.49	4.97	5.08	2.52	0.97	0.09
	max	0.12	1.07	2.59	5.06	5.15	2.59	1.03	0.10
0.25	min	0.94	1.00	2.51	5.05	5.08	2.53	1.00	0.09
	max	0.11	1.08	2.57	5.09	5.14	2.60	1.03	0.10

k = 10, n = 5

		α(%) lower tail				α(%) upper tail			
r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
2.0	min	0.08	0.97	2.39	4.89	4.76	2.34	0.97	0.08
	max	0.13	1.02	2.63	5.08	5.13	2.63	1.08	0.11
1.5	min	0.09	0.99	2.47	4.98	4.82	2.39	0.93	0.08
	max	0.12	1.07	2.60	5.09	5.07	2.56	1.03	0.11
1.0	min	0.09	1.01	2.49	4.94	4.91	2.44	0.95	0.09
	max	0.13	1.07	2.60	5.06	5.01	2.54	1.00	0.10
0.5	min	0.10	0.98	2.48	4.96	4.94	2.44	0.96	0.09
	max	0.12	1.06	2.60	5.08	5.00	2.48	1.00	0.10
0.25	min	0.10	0.98	2.50	4.99	4.91	2.42	0.96	0.09
	max	0.12	1.02	2.58	5.07	5.00	2.46	0.99	0.10

Table B.14

Minimum and maximum values of $\alpha_R(\%)$ found for the 11 values of c, in a One and Two-way Analysis of Variance for k = 3 and n = 5

			$\alpha(\%)$ lower tail				$\alpha(\%)$ upper tail			
	r		0.1	1.0	2.5	5.0	5.0	2.5	1.0	0.1
One-way	1.5	min	3.42	3.42	3.42	3.49	4.67	2.51	0.87	0.08
		max	3.56	3.56	3.56	3.63	4.85	2.79	0.93	0.16
	1.0	min	1.66	1.63	1.69	4.39	4.66	2.45	1.00	0.10
		max	1.77	1.77	1.80	4.57	4.82	2.54	1.04	0.12
	0.5	min	0.43	0.62	2.65	4.99	4.91	2.48	0.97	0.10
		max	0.49	0.69	2.79	5.08	5.06	2.56	1.06	0.12
Two-way	1.5	min	3.39	3.39	3.39	3.63	4.73	2.27	0.92	0.09
		max	3.49	3.49	3.49	3.73	4.88	2.36	0.97	0.13
	1.0	min	1.62	1.62	1.76	4.51	4.85	2.40	1.00	0.09
		max	1.74	1.74	1.87	4.68	5.00	2.50	1.05	0.12
	0.5	min	0.43	0.70	2.50	4.88	4.89	2.45	0.98	0.10
		max	0.48	0.76	2.60	4.96	5.02	2.58	1.02	0.12

APPENDIX C : COMPUTER PROGRAMS AND OUTPUT FOR CHAPTER 5

In this appendix computer programs written to find the significance and power level of a test for rounded Johnson data are considered. A list of output produced by these programs is provided. The appendix also contains tables of results for α_{JR} (significance level of a test under rounding) and P_{FR} (power of a test under rounding) which are referred to in Chapter 5. Finally details of the method used in section (5.3.1) for obtaining the mean and variance of the test statistic are given.

The FORTRAN programs SIMUL, PSIMUL referred to in Appendix B were modified to allow samples to be drawn from Johnson distributions with shape parameters $\sqrt{\beta_1}$ and β_2. These two programs estimated the significance and power level of a test for Johnson rounded data by Monte Carlo methods. A program USIMUL was also written to obtain the significance and power level of a list for Johnson distributions whch are not subject to rounding.

The following is a list of all the output produced by the programs SIMUL, PSIMUL and USIMUL, upon which the study of the effect of rounded non-normal data on the significance and power level of a test was based.

Significance level of a test

The significance level of each test for unrounded data (α_J) and rounded data (α_{JR}) from 29 Johnson distributions with shape parameters ($\sqrt{\beta_1}$,β_2) were evaluated for values corresponding to the lower and upper 5% points under normal theory conditions. The shape parameters fell on a grid with $\sqrt{\beta_1}$ = 0.0, 0.2, 0.4, 0.6

and 0.8 and β_2 = 2.0, 2.4, 2.8, 3.2, 3.6 and 4.4. Omitting β_2 = 2.0 and $\sqrt{\beta_1}$ = 0.8 produced 29 distributions. For each of the 29 distributions

(i) α_J was obtained for combinations of n and k given below

(ii) α_{JR} was obtained for 11 values of c(−0.5,...,0.5), for combinations of n, k and r given below.

The results from SIMUL and USIMUL programs were based on 10,000 iterations.

<u>One sample t−test</u>

α_J: n = 10, 25

α_{JR}: n = 10, 25 with r = 0.5, 1.0, 1.5

<u>Chi−squared test</u>

α_J: n = 10, 25

α_{JR}: n = 10, 25 with r = 0.25, 0.5, 1.0

<u>Two sample t−test</u>

α_J: n = 10, 25

α_{JR}: n = 10, 25 with r = 1.0, 1.5, 2.0

<u>F−test</u>

α_J: n = 10, 25

α_{JR}: n = 10, 25 with r = 0.5, 1.0, 1.5

One way analysis of variance

α_J: k = 3, n = 10, 25

k = 5, n = 10, 25

α_{JR}: k = 3, n = 10, 25 / k = 5, n = 10, 25 } with r = 0.5, 1.0, 1.5, 2.0

Two way analysis of variance

α_J: k = 3, n = 10, 25

α_{JR}: k = 3, n = 10, 25 with r = 0.5, 1.0, 1.5, 2.0

A listing of all output mentioned above is available on request.

Power of a test

The power level of each test for unrounded data (P_J) and rounded data (P_{JR}) from a Johnson distribution with shape parameters ($\sqrt{\beta_1}$,β_2) was found. The power was evaluated for values of the alternative hypothesis H_1 (one tailed) corresponding to powers of 0.3 and 0.7 under normal theory conditions, where $\alpha = 0.05$. For Johnson distributions corresponding to a kurtosis of β_2 given below

(i) P_J was obtained for values of n stated below.

(ii) P_{JR} was obtained for 11 values of c(−0.5,...,0.5) for combinations of n, k and r stated below.

One sample t-test

P_J: n = 10, 25

P_{JR}: n = 10, 25 with r = 0.5, 1.0, 1.5 } for β_2 = 2.0, 2.4, 4.4

Chi-squared test

P_J: n = 10, 25

P_{JR}: n = 10, 25 with r = 0.25, 0.5, 1.0

} for β_2 = 2.0, 2.4, 3.6, 4.4

Two sample t-test

P_J: n = 10, 25

P_{JR}: n = 10, 25 with r = 1.0, 1.5, 2.0

} for β_2 = 2.0, 2.4, 2.8, 3.2, 3.6, 4.4

F-test

P_J: n = 10, 25

P_{JR}: n = 10, 25 with r = 0.5, 1.0, 1.5

} for β_2 = 2.4, 2.8, 3.2, 3.6

One way analysis of variance

α_J: k = 3, n = 10, 25

k = 5, n = 10, 25

α_{JR}: k = 3, n = 10, 25

k = 5, n = 10, 25 with r = 1.0, 1.5, 2.0

} for β_2 = 2.0, 2.4, 4.4

Two way analysis of variance

α_J: k = 3, n = 10

α_{JR}: k = 3, n = 10

} for β_2 = 2.0 and 4.4 each at $\sqrt{\beta_1}$ = 0.0 and 0.8

A list of all output mentioned above is available on request.

Table C.1

Range of $\alpha_{JR}(\%)$ values found for 11 values of c, in a one sample t-test for n = 10 and α = 0.05, where r = 1.5, 1.0 and 0.5. J gives the level of significance ($\alpha_J\%$) where the Johnson distribution is not subject to rounding.

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
	J	5.1	5.1	5.8	4.4	6.7	3.8	7.8	3.3		
2.0	0.5	5.0-5.3	5.0-5.3	5.7-6.0	4.4-4.6	6.6-7.0	3.8-4.1	7.2-8.2	3.1-3.7	———	
	1.0	4.8-5.5	4.8-5.5	5.5-6.3	4.1-4.9	5.9-7.5	3.4-4.6	5.4-9.3	2.7-4.5		
	1.5	3.5-7.2	3.5-7.3	4.1-8.3	3.2-6.6	4.2-9.7	3.1-6.4	4.4-11.6	2.4-9.9		
	J	5.1	5.1	5.7	4.4	6.6	3.8	7.7	3.3	8.9	2.8
2.4	0.5	5.0-5.3	5.0-5.3	5.7-6.0	4.4-4.6	6.6-6.8	3.8-4.0	7.5-7.9	3.3-3.5	8.3-9.2	2.7-3.1
	1.0	4.8-5.6	4.8-5.6	5.4-6.3	4.3-4.9	6.2-7.3	3.8-4.3	6.5-8.6	3.2-3.9	6.7-10.8	2.3-4.9
	1.5	3.5-7.1	3.6-7.2	3.9-8.0	3.3-6.5	4.5-9.3	3.1-5.8	4.6-11.0	3.1-5.9	3.2-13.7	2.0-7.5
	J	5.1	5.1	5.7	4.5	6.5	3.9	7.5	3.4	8.7	2.9
2.8	0.5	5.0-5.3	5.0-5.2	5.6-5.9	4.4-4.5	6.4-6.6	3.9-4.0	7.3-7.6	3.3-3.5	8.5-8.8	2.9-3.0
	1.0	4.8-5.6	4.8-5.6	5.4-6.3	4.3-5.0	6.1-7.0	3.8-4.4	6.8-8.2	3.4-3.8	6.4-9.8	2.8-3.4
	1.5	3.4-7.0	3.5-7.0	3.7-7.0	3.2-6.3	4.2-8.8	3.1-5.7	5.0-10.3	3.0-5.4	5.3-12.5	2.5-5.8
	J	5.0	5.0	5.6	4.5	6.5	3.8	7.3	3.5	8.4	2.9
3.2	0.5	5.0-5.2	5.0-5.2	5.5-5.8	4.5-4.6	6.4-6.6	3.8-4.0	7.1-7.3	3.4-3.6	8.2-8.5	2.9-3.1
	1.0	4.9-5.7	4.8-5.6	5.4-6.3	4.4-5.1	6.1-7.0	3.8-4.4	6.7-7.9	3.4-3.9	7.2-9.2	2.9-3.4
	1.5	3.3-6.8	3.3-6.8	3.6-7.6	3.1-6.2	4.0-8.6	2.9-5.6	4.5-9.7	2.8-5.2	5.6-11.5	2.8-5.0
	J	5.0	5.0	5.5	4.5	6.2	4.0	7.1	3.5	8.2	3.0
3.6	0.5	4.9-5.2	4.9-5.2	5.5-5.7	4.4-4.6	6.1-6.4	3.9-4.2	6.9-7.3	3.4-3.7	8.0-8.2	3.0-3.1
	1.0	4.9-5.7	4.8-5.7	5.3-6.3	4.3-5.1	5.8-6.9	3.9-4.6	6.6-7.9	3.4-4.0	7.2-8.8	3.0-3.5
	1.5	3.1-6.9	3.2-7.0	3.4-7.4	3.0-6.5	3.8-8.2	2.9-6.0	4.4-9.6	2.8-5.4	5.1-10.9	2.7-4.7
	J	5.0	4.9	5.4	4.5	6.0	4.0	6.8	3.6	7.6	3.2
4.4	0.5	4.9-5.2	4.9-5.1	5.4-5.7	4.4-4.6	5.9-6.2	4.0-4.2	6.7-6.8	3.6-3.7	7.5-7.7	3.2-3.3
	1.0	4.7-5.7	4.7-5.7	5.1-6.2	4.4-5.2	5.6-6.8	4.0-4.7	6.1-7.5	3.6-4.2	6.7-8.3	3.2-3.7
	1.5	2.9-7.3	3.0-7.4	3.1-7.7	2.8-6.9	3.4-8.0	2.7-6.4	3.8-8.7	2.6-5.9	4.4-9.8	2.5-5.2

Table C.2

Range of $\alpha_{JR}(\%)$ values found for 11 values of c, in a one sample t-test for n = 25 and α = 0.05, where r = 1.5, 1.0 and 0.5. J gives the level of significance ($\alpha_J\%$) where the Johnson distribution is not subject to rounding.

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
	J	5.4	5.5	5.8	4.9	6.4	4.3	6.9	3.8		
2.0	0.5	5.2-5.5	5.3-5.6	5.6-6.0	4.7-5.1	5.9-6.6	4.1-4.8	5.2-7.6	3.3-5.2		
	1.0	5.0-5.8	5.0-5.8	5.2-6.3	4.5-5.4	4.8-7.2	3.8-6.0	3.3-9.7	2.4-8.3		
	1.5	4.3-6.5	4.6-6.6	4.3-6.9	4.0-6.7	3.2-9.1	2.9-8.5	1.6-12.6	1.7-16.6		
	J	5.4	5.4	5.9	4.9	6.4	4.3	6.9	3.9	7.4	3.4
2.4	0.5	5.2-5.5	5.3-5.5	5.7-6.0	4.8-5.1	6.1-6.5	4.3-4.6	3.8-4.2	5.8-7.6	5.8-7.6	2.9-4.4
	1.0	5.2-5.8	5.0-5.6	5.5-5.9	4.6-5.2	5.9-6.5	4.3-4.7	5.5-7.6	3.4-5.0	3.1-10.4	2.0-9.1
	1.5	4.7-5.8	4.9-5.9	5.1-6.2	4.6-5.6	5.1-7.4	3.9-5.5	4.1-10.0	2.9-8.5	1.7-15.1	1.4-14.8
	J	5.4	5.3	5.9	4.9	6.3	4.4	7.0	4.0	7.5	3.5
2.8	0.5	5.2-5.5	5.2-5.6	5.7-6.0	4.8-5.1	6.2-6.5	4.3-4.5	6.6-6.8	3.9-4.3	7.0-7.5	3.4-3.9
	1.0	5.2-5.5	5.0-5.7	5.4-6.1	4.7-5.2	5.8-6.4	4.2-4.9	6.4-6.9	3.9-4.4	5.4-8.6	3.0-4.9
	1.5	4.8-5.8	4.8-5.8	5.4-6.3	4.7-5.3	5.6-6.6	4.1-5.1	5.4-7.9	3.7-5.6	3.9-10.6	2.5-8.9
	J	5.4	5.4	5.9	4.9	6.3	4.3	6.8	4.0	7.4	3.6
3.2	0.5	5.2-5.5	5.1-5.4	5.6-5.9	4.7-5.0	6.2-6.3	4.2-4.4	6.6-6.8	3.9-4.2	7.0-7.3	3.5-3.9
	1.0	5.1-5.6	4.9-5.4	5.4-6.0	4.7-5.1	5.9-6.4	4.2-4.7	6.3-6.7	3.9-4.5	6.5-7.5	3.5-4.1
	1.5	5.0-5.8	4.8-5.9	5.1-6.1	4.5-5.5	5.7-6.8	3.9-4.8	5.9-7.0	3.8-4.9	5.3-8.1	3.5-4.6
	J	5.4	5.3	5.9	4.9	6.3	4.4	6.8	4.0	7.3	3.6
3.6	0.5	5.2-5.5	5.2-5.3	5.6-5.9	4.8-5.0	6.2-6.3	4.4-4.5	6.7-6.9	3.9-4.2	7.2-7.4	3.5-3.8
	1.0	5.0-5.5	5.1-5.5	5.3-6.1	4.7-5.2	5.6-6.3	4.3-4.9	6.3-6.9	3.8-4.4	6.9-7.4	3.4-4.1
	1.5	5.0-5.9	4.7-5.9	5.3-6.2	4.3-5.6	5.6-6.7	3.9-5.1	6.0-7.3	3.8-4.9	6.3-7.9	3.5-4.7
	J	5.4	5.3	5.8	4.8	6.2	4.4	6.5	4.0	7.0	3.7
4.4	0.5	5.2-5.5	5.0-5.4	5.5-5.9	4.7-5.0	6.1-6.2	4.3-4.6	6.4-6.7	4.0-4.2	7.0-7.2	3.6-3.8
	1.0	4.9-5.5	5.0-5.6	5.3-5.9	4.7-5.2	5.5-6.4	4.3-4.8	6.0-6.8	3.9-4.5	6.7-7.4	3.7-4.2
	1.5	4.7-6.2	4.3-6.2	5.1-6.5	4.0-5.8	5.5-6.7	3.8-5.3	5.9-7.3	3.7-4.8	6.4-7.6	3.5-4.5

Table C.3

Range of P_{JR} (%) values found for 11 values of c, in a one sample t-test for n = 10 and P = 0.30 distribution is not subject to rounding. J gives the level of power (P_J%) where the Johnson and 0.70, where r = 1.5, 1.0 and 0.5.

P = 0.30

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
	J	27.8	27.7	29.0	26.2	30.0	24.6	30.8	22.8		
2.0	0.5	27.4-27.9	27.2-27.9	28.7-29.0	25.9-26.4	29.6-30.2	23.8-25.4	29.3-31.5	20.9-26.0		
	1.0	26.2-28.2	26.4-28.1	26.3-29.9	24.1-26.7	26.7-31.4	21.5-28.0	27.3-34.1	17.9-35.2	———	
	1.5	22.3-32.0	22.2-31.6	21.9-31.8	21.3-32.9	21.2-31.7	17.9-32.6	22.3-35.0	13.2-57.2		
	J	28.8	28.7	30.0	27.3	31.0	25.7	31.8	23.8	32.3	21.7
2.4	0.5	28.4-28.9	28.2-28.9	29.6-30.1	26.8-27.5	30.5-31.1	25.3-25.9	31.1-31.9	23.2-24.3	31.2-32.7	20.0-25.4
	1.0	26.9-28.8	26.4-28.6	27.6-29.9	25.3-27.4	28.3-31.0	24.0-26.0	28.9-32.0	21.5-26.2	27.9-34.2	16.5-39.1
	1.5	25.2-31.6	22.5-31.2	25.9-32.4	21.8-30.8	26.1-33.0	21.4-31.1	23.3-34.4	17.7-32.3	24.5-38.8	12.0-55.0
	J	32.0	31.9	32.9	30.8	33.8	29.6	34.6	28.1	35.3	26.4
4.4	0.5	31.3-32.0	31.3-31.8	32.2-33.0	30.2-30.7	33.2-33.9	28.9-29.5	33.9-34.8	27.4-28.2	34.7-35.4	25.8-26.5
	1.0	29.4-31.8	29.4-31.8	30.3-32.7	28.4-30.7	31.1-33.4	27.3-29.4	31.9-34.2	26.0-28.3	32.5-35.0	24.5-26.8
	1.5	27.7-33.4	27.9-33.0	23.7-34.5	26.8-32.0	24.7-35.5	20.7-31.4	29.3-36.5	23.9-30.8	26.6-34.7	21.4-30.0

P = 0.70

	J	68.6	68.3	67.9	69.2	67.1	70.2	66.2	71.6		
2.0	0.5	67.7-68.3	67.4-67.9	66.8-67.7	67.9-69.2	66.1-66.8	68.4-70.5	64.2-67.1	66.2-80.5		
	1.0	63.9-67.9	63.6-67.2	63.7-66.6	64.1-68.8	62.0-66.8	62.6-77.7	60.5-68.2	55.6-97.4	———	
	1.5	57.9-66.0	57.7-71.0	58.9-69.5	56.5-72.1	55.3-65.0	51.7-83.5	49.7-69.3	41.5-99.0		
	J	69.2	68.3	68.5	69.7	67.7	70.6	66.9	71.9	66.0	73.6
2.4	0.5	68.2-69.0	68.0-68.7	67.6-68.4	68.6-69.4	67.0-67.5	69.5-70.2	66.1-66.2	70.0-72.0	64.9-66.0	67.8-82.0
	1.0	65.4-67.8	65.1-67.4	65.0-67.3	65.6-66.9	64.5-65.9	65.6-69.5	63.3-66.9	62.1-76.9	59.0-68.6	55.4-96.2
	1.5	61.4-68.4	59.8-68.0	59.8-67.1	60.4-69.4	59.2-66.1	57.4-70.5	57.0-66.2	52.2-87.2	48.5-66.9	40.6-99.8
	J	71.8	71.4	71.3	71.9	70.8	72.4	70.3	72.9	69.6	73.6
4.4	0.5	70.9-71.3	70.6-71.1	70.4-70.9	70.9-71.3	70.0-70.6	71.4-71.9	69.4-70.0	71.8-72.3	68.8-69.4	72.4-72.9
	1.0	67.9-69.9	67.7-69.4	67.7-69.6	68.0-69.8	67.4-69.2	68.4-70.2	67.0-68.9	68.5-70.4	66.5-68.5	68.8-71.2
	1.5	61.5-69.9	61.2-69.6	61.9-69.7	60.7-69.6	62.1-69.4	60.2-69.6	62.0-68.8	60.6-70.3	61.7-68.0	59.0-72.0

Table C4

Range of $\alpha_{JR}(\%)$ values found for 11 values of c, in a chi-squared test for a variance for n = 10 and α = 0.05, where r = 1.0, 0.5 and 0.25. J gives the level of significance ($\alpha_J\%$) where the Johnson distribution is not subject to rounding.

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
2.0	J	2.1	1.1	2.3	1.0	2.8	0.8	4.0	0.5		
	0.25	2.1-2.1	1.1-1.2	2.2-2.3	1.0-1.1	2.7-2.8	0.9-1.0	3.8-4.3	0.3-0.8		
	0.5	2.0-2.0	1.4-1.8	2.1-2.2	1.3-1.7	2.5-2.8	1.0-1.6	3.1-4.7	0.4-1.5	—	
	1.0	1.4-1.5	1.2-4.3	1.4-1.7	0.9-4.5	1.4-2.8	0.9-5.0	1.5-5.5	0.6-3.5		
2.4	J	3.3	3.0	3.3	3.0	3.5	2.5	4.4	2.2	6.3	1.5
	0.25	3.2-3.3	2.8-2.9	3.2-3.3	2.8-2.9	3.4-3.5	2.6-2.7	4.3-4.4	2.2-2.3	6.0-6.5	1.3-1.8
	0.5	3.0-3.1	3.3-3.3	3.0-3.1	3.2-3.3	3.2-3.4	3.0-3.2	3.9-4.3	2.5-3.1	5.0-7.0	0;7-3.2
	1.0	2.2-2.3	4.4-4.9	2.2-2.3	4.3-4.8	2.3-2.5	3.9-5.0	2.3-4.1	1.8-6.4	2.3-8.8	0.9-8.2
2.8	J	4.4	4.4	4.4	4.3	4.5	4.3	4.9	4.1	6.5	3.6
	0.25	4.3-4.4	4.5-4.6	4.3-4.4	4.4-4.5	4.4-4.5	4.4-4.5	4.8-4.9	4.2-4.3	6.3-6.5	3.7-3.9
	0.5	4.1-4.2	4.9-4.0	4.0-4.2	4.9-4.9	4.1-4.2	4.9-5.0	4.5-4.6	4.7-4.9	5.6-6.2	3.9-4.6
	1.0	2.8-3.0	6.1-6.3	2.9-3.0	6.0-6.2	2.9-3.0	6.0-6.3	3.1-3.5	5.4-6.7	3.0-6.4	3.6-7.6
3.2	J	5.5	5.7	5.4	5.7	5.4	5.6	5.5	5.7	6.6	5.5
	0.25	5.3-5.4	5.7-5.8	5.3-5.4	5.6-5.7	5.2-5.3	5.6-5.7	5.2-5.3	5.8-5.9	6.4-6.5	5.7-5.8
	0.5	5.0-5.1	6.2-6.3	5.0-5.0	6.1-6.2	4.9-5.0	6.1-6.2	5.0-5.2	6.2-6.3	6.0-6.2	6.1-6.3
	1.0	3.5-3.7	7.3-7.5	3.5-3.7	7.2-7.4	3.4-3.6	7.2-7.4	3.5-3.7	7.2-7.6	3.6-4.9	6.4-8.4
3.6	J	6.4	6.7	6.3	6.7	6.2	6.6	6.4	6.7	6.8	6.8
	0.25	6.2-6.3	6.7-6.8	6.1-6.2	6.7-6.8	6.1-6.2	6.5-6.6	6.3-6.4	6.7-6.8	6.7-6.8	6.8-6.9
	0.5	5.8-5.9	6.8-6.9	5.8-5.9	6.8-6.9	5.7-5.8	6.9-7.0	6.0-6.1	7.0-7.2	6.2-6.4	7.3-7.4
	1.0	4.1-4.2	8.1-8.4	4.1-4.2	8.1-8.3	3.9-4.2	8.0-8.2	4.0-4.4	8.2-8.7	4.3-4.7	8.1-8.8
4.4	J	7.9	7.9	7.9	7.8	7.8	7.7	7.7	7.7	7.6	7.7
	0.25	7.7-7.8	8.0-8.1	7.7-7.8	7.9-8.0	7.7-7.8	7.8-7.9	7.5-7.6	7.7-7.8	7.4-7.6	7.7-7.8
	0.5	7.2-7.3	8.4-8.4	7.2-7.3	8.3-8.4	7.1-7.2	8.2-8.3	7.0-7.1	8.1-8.2	7.0-7.1	8.1-8.2
	1.0	5.1-5.2	9.2-9.4	5.1-5.2	9.1-9.3	5.0-5.2	9.1-9.2	4.9-5.1	9.0-9.1	4.8-5.1	9.1-9.2

Table C.5

Range of $\alpha_{JR}(\%)$ values found for 11 values of c in a chi-squared test for a variance for n = 25 and α = 0.05, where r = 1.0, 0.5 and 0.25. J gives the level of significance ($\alpha_J\%$) where the Johnson distribution is not subject to rounding.

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
		lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
2.0	J	1.5	1.0	1.6	1.0	1.8	0.8	1.8	0.6		
	0.25	1.4-1.5	1.1-1.2	1.5-1.6	1.0-1.1	1.8-1.8	0.8-0.9	2.1-2.7	0.3-0.8		
	0.5	1.3-1.4	1.4-1.7	1.4-1.5	1.3-1.7	1.5-1.9	1.0-1.8	1.6-3.5	0.4-1.9		
	1.0	0.9-1.1	1.5-6.7	0.8-1.3	1.1-7.0	0.7-2.3	1.2-7.6	0.5-5.8	0.6-6.2		
2.4	J	2.9	2.4	3.1	2.5	3.1	2.3	3.5	2.1	4.3	1.6
	0.25	2.8-3.0	2.5-2.7	2.8-3.0	2.5-2.7	2.9-3.2	2.4-2.6	3.3-3.6	2.2-2.4	3.9-4.6	1.4-2.2
	0.5	2.6-2.8	3.2-3.4	2.5-2.8	3.2-3.4	2.6-2.9	3.0-3.4	3.0-3.4	2.9-3.6	3.1-5.7	0.6-4.2
	1.0	1.7-2.0	5.9-6.6	1.7-2.0	5.7-6.7	1.7-2.3	4.9-7.2	1.5-3.8	2.0-9.6	1.0-9.5	0.7-13.2
2.8	J	4.5	4.1	4.5	4.0	4.5	4.2	4.8	3.9	5.7	3.5
	0.25	4.3-4.4	4.1-4.4	4.3-4.5	4.1-4.3	4.3-4.5	4.2-4.5	4.5-4.8	4.0-4.3	5.4-5.7	3.6-4.0
	0.5	3.8-4.1	4.8-5.1	3.8-4.1	4.8-5.1	3.8-4.2	4.9-5.2	4.1-4.4	4.7-5.1	4.3-5.5	4.0-5.1
	1.0	2.5-2.9	7.7-8.5	2.5-2.9	7.6-8.3	2.5-2.9	7.6-8.5	2.6-3.4	6.8-9.0	2.2-6.2	3.8-11.2
3.2	J	5.9	5.6	5.9	5.5	5.9	5.7	6.1	5.8	6.9	5.4
	0.25	5.6-5.8	5.8-6.0	5.6-5.9	5.7-5.9	5.7-5.9	5.7-5.9	5.7-5.9	5.7-6.0	6.3-6.7	5.6-5.8
	0.5	5.1-5.3	6.5-6.7	5.1-5.3	6.4-6.7	5.0-5.4	6.4-6.7	5.2-5.5	6.6-6.7	5.7-6.3	6.3-6.8
	1.0	3.3-3.7	9.1-9.9	3.3-3.7	9.3-9.8	3.3-3.7	9.4-9.9	3.5-3.8	9.2-10.2	3.4-5.1	8.3-11.1
3.6	J	7.2	6.9	7.3	6.9	7.3	6.8	7.5	6.9	7.8	7.1
	0.25	6.8-7.0	7.1-7.3	6.9-7.1	7.1-7.2	6.8-7.0	6.9-7.2	7.0-7.4	7.0-7.2	7.4-7.6	7.2-7.4
	0.5	6.1-6.5	7.9-8.1	6.2-6.5	7.7-8.0	6.1-6.4	7.6-7.9	6.3-6.6	7.7-8.0	6.6-7.0	7.9-8.3
	1.0	4.0-4.5	10.2-11.1	4.0-4.5	10.5-11.2	4.1-4.5	10.2-10.9	4.2-4.5	10.3-11.0	4.3-5.0	10.5-11.5
4.4	J	9.5	8.8	9.4	8.7	9.4	8.6	9.3	8.5	9.2	8.7
	0.25	9.0-9.3	8.8-9.1	9.0-9.4	8.8-9.1	8.7-9.2	8.8-8.9	8.9-9.1	8.6-8.9	8.8-9.1	8.7-9.1
	0.5	8.2-8.6	9.4-9.8	8.1-8.5	9.4-9.9	8.1-8.4	9.5-9.7	8.1-8.4	9.3-9.6	8.0-8.3	9.5-9.8
	1.0	5.3-6.0	12.0-12.7	5.4-6.0	11.8-12.8	5.4-6.1	12.0-12.6	5.3-5.9	11.9-12.5	5.3-5.9	11.9-12.5

Table C.6

Range of $P_{JR}(\%)$ values found for 11 values of c, in a chi-squared test for a variance for n = 10 and P = 0.3 and 0.7, where r = 0.25, 0.5 and 0.25. J gives the level of power ($P_J\%$) where the Johnson distribution is not subject to rounding.

P = 0.30

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
	J	25.4	29.8	25.4	29.9	25.5	30.3	25.5	31.3	24.7	32.7
2.4	0.25	24.9-25.0	30.1-30.2	24.9-25.0	30.2-30.3	25.1-25.2	30.5-30.6	24.8-25.4	31.4-31.5	23.3-25.8	32.1-33.5
	0.5	23.0-23.2	31.0-31.2	23.0-23.1	31.2-31.3	23.1-23.6	31.4-31.5	22.1-25.0	31.8-33.0	19.5-27.9	30.9-36.3
	1.0	15.5-16.9	32.5-32.7	15.2-23.1	32.5-32.8	13.8-19.5	32.1-33.9	11.4-26.2	29.8-36.7	8.7-35.8	25.7-40.3
	J	32.8	28.7	32.7	28.9	32.8	28.9	33.0	29.4	33.2	29.6
3.6	0.25	31.9-32.4	28.6-28.9	31.7-32.4	28.9-29.1	31.7-32.3	28.8-29.3	32.0-32.6	29.4-30.1	32.3-33.0	29.7-29.9
	0.5	29.8-30.2	29.4-30.1	29.6-30.1	29.7-30.2	29.5-30.2	29.7-30.4	29.1-30.0	29.9-30.5	29.8-30.9	30.4-30.7
	1.0	21.2-21.8	31.0-31.7	21.2-21.9	31.0-31.6	20.8-22.2	31.1-31.5	21.0-24.1	31.0-31.8	18.6-24.9	31.6-32.4

P = 0.70

	J	70.2	74.6	70.1	74.7	69.7	74.5	68.7	74.5	67.3	74.3
2.4	0.25	68.6-68.9	74.6-74.7	68.6-68.7	74.6-74.7	68.2-68.4	74.4-74.6	66.7-68.2	74.5-74.6	63.1-69.7	69.8-78.4
	0.5	63.7-64.1	75.0-75.1	63.7-64.2	75.0-75.1	62.5-64.6	74.8-75.0	59.0-67.1	74.7-75.0	50.7-69.9	74.4-76.4
	1.0	40.9-53.6	74.8-75.2	39.8-55.1	74.8-75.1	35.7-61.4	74.6-75.1	29.6-75.6	73.5-75.6	21.5-93.6	75.0-75.6
	J	71.3	67.2	71.1	67.3	71.1	67.3	70.9	68.1	70.4	66.8
3.6	0.25	69.9-70.2	67.2-67.5	69.7-70.2	67.2-67.8	69.7-70.1	67.2-67.6	69.4-70.0	68.1-68.4	69.2-69.6	66.7-67.0
	0.5	65.8-66.4	67.5-68.3	65.7-66.4	67.5-68.4	65.8-66.4	67.7-68.1	66.0-66.8	68.8-69.0	64.9-65.9	67.1-67.7
	1.0	49.0-52.0	67.9-68.5	49.0-51.5	67.8-68.5	47.8-51.5	67.9-68.6	47.1-55.6	69.1-69.8	41.6-59.1	67.3-68.0

Table C7

Lower tail range of $\alpha_{JR}(\%)$ values found for 11 values of c in the two sample t-test for n = 10 and α = 0.05, where r = 2.0, 1.5 and 1.0. J gives the level of significance ($\alpha_J\%$) when the Johnson distribution is not subject to rounding.

		0.0	0.2	0.4	0.6	0.8
β_2	r	lower tail	lower tail	lower tail	lower tail	lower tail
	J	5.1	5.1	5.1	5.1	
2.0	1.0	5.0-5.1	5.0-5.2	5.0-5.2	5.0-5.1	
	1.5	4.9-5.1	4.8-5.2	4.7-5.2	4.7-5.1	———
	2.0	4.9-5.7	4.9-5.5	4.8-5.5	4.8-5.4	
	J	5.1	5.1	5.1	5.1	5.1
2.4	1.0	5.0-5.2	5.0-5.1	5.0-5.1	5.0-5.2	4.9-5.1
	1.5	4.9-5.1	4.8-5.2	4.8-5.1	4.7-5.2	4.7-5.1
	2.0	4.9-5.2	4.9-5.3	4.9-5.2	4.8-5.2	4.8-5.1
	J	5.1	5.1	5.1	5.1	5.0
2.8	1.0	5.0-5.2	5.0-5.1	5.0-5.1	5.0-5.1	5.0-5.1
	1.5	4.9-5.1	4.8-5.1	4.8-5.1	4.9-5.1	4.7-5.1
	2.0	4.9-5.2	4.9-5.2	4.9-5.2	4.9-5.2	4.7-5.2
	J	5.1	5.1	5.1	5.1	5.1
3.2	1.0	5.0-5.1	5.0-5.1	5.0-5.1	5.0-5.1	5.0-5.1
	1.5	4.9-5.1	4.8-5.1	4.8-5.2	4.8-5.1	4.7-5.1
	2.0	4.9-5.2	4.9-5.1	4.9-5.2	4.9-5.2	4.7-5.2
	J	5.1	5.1	5.1	5.0	5.0
3.6	1.0	5.0-5.1	5.0-5.1	5.0-5.1	5.0-5.1	5.0-5.1
	1.5	4.9-5.1	4.9-5.1	4.8-5.1	4.8-5.1	4.8-5.1
	2.0	4.9-5.2	4.9-5.2	4.9-5.2	4.9-5.1	4.9-5.1
	J	5.0	5.1	5.1	5.0	5.0
4.4	1,0	5.0-5.1	5.0-5.1	5.0-5.1	5.0-5.1	5.0-5.1
	1.5	4.9-5.1	4.9-5.1	4.8-5.1	4.8-5.2	4.8-5.1
	2.0	4.9-5.2	4.9-5.2	4.9-5.2	5.0-5.3	4.9-5.2

Table C.8

Lower tail range of the significance levels ($\alpha_R(\%)$ found for 11 values of c in the two sample t–test for n = 10 and α = 0.05 for rounded normal data.

r	lower tail
1.0	5.1 - 5.2
1.5	4.9 - 5.1
2.0	4.9 - 5.3

Table C.9

Range of P_{JR}(%) values found for 11 values of c, in a two sample t-test for n = 10 and P = 0.3 and 0.7, where r = 1.0, 1.5 and 2.0. J gives the level of power (P_J%) where the Johnson distribution is not subject to rounding.

P = 0.30

	$\sqrt{\beta_1}$	0.0	0.2	0.4	0.6	0.8
β_2	T	lower tail	lower tail	lower tail	lower tail	lower tail
2.0	J	29.1	28.9	28.8	29.5	
	1.0	26.5-29.0	25.9-29.6	24.7-33.0	19.9-37.4	
	1.5	23.8-28.5	23.3-30.3	20.3-37.0	15.7-45.7	———
	2.0	19.9-27.9	19.9-27.9	17.1-33.1	12.2-49.7	
2.4	J	29.8	29.5	29.5	29.4	30.0
	1.0	27.7-28.6	27.7-28.7	27.3-28.9	25.7-31.7	19.1-41.3
	1.5	26.6-26.9	25.9-27.2	24.8-28.8	20.5-34.5	14.8-53.5
	2.0	24.0-25.3	23.6-25.7	21.2-27.6	17.4-34.7	11.9-55.8
4.4	J	31.2	31.1	31.1	31.1	31.0
	1.0	29.1-30.4	29.2-30.4	29.4-30.2	29.3-29.9	29.4-29.9
	1.5	26.9-28.3	26.7-28.5	27.3-28.7	27.2-28.7	27.5-28.4
	2.0	23.4-27.9	23.4-27.7	23.7-27.5	24.1-27.7	23.8-27.8

P = 0.70

2.0	J	70.0	70.4	70.5	70.4	
	1.0	66.0-68.4	66.0-68.7	65.6-69.6	63.9-72.8	
	1.5	59.9-67.7	59.2-69.3	54.5-73.8	44.1-77.4	———
	2.0	51.8-64.3	49.9-67.8	42.8-76.2	28.9-88.5	
2.4	J	70.4	70.5	70.7	70.5	70.3
	1.0	67.1-67.9	67.1-67.8	66.9-68.3	65.0-70.7	63.1-73.8
	1.5	63.0-64.6	62.5-64.8	60.9-68.3	55.9-74.4	43.1-82.4
	2.0	58.0-60.2	56.8-61.2	52.8-65.6	46.2-75.3	30.2-84.9
4.4	J	72.1	72.0	71.9	72.1	71.8
	1.0	68.6-69.3	68.7-69.1	68.3-69.0	68.4-69.0	68.2-69.1
	1.5	64.1-64.9	64.1-65.8	64.2-66.0	64.5-65.8	64.3-66.1
	2.0	55.9-64.4	55.9-64.4	56.0-63.9	57.0-64.1	56.4-64.6

Table C.10

Range of $\alpha_{JR}(\%)$ values found for 11 values of c in an F-test for equality of two variances for n = 10 and α = 0.05, where r = 1.5, 1.0 and 0.5. J gives the level of significance ($\alpha_J\%$) where the Johnson distribution is not subject to rounding.

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
	J	2.0	2.0	2.1	2.1	2.5	2.5	3.3	3.4		
2.0	0.5	1.9-2.1	2.0-2.1	2.0-2.2	2.0-2.2	2.4-2.5	2.4-2.5	2.9-3.8	2.9-3.8		
	1.0	1.9-2.7	2.0-2.7	2.0-2.8	1.9-2.9	2.1-3.3	2.0-3.4	2.2-5.2	2.1-5.2	-	
	1.5	1.8-3.7	1.8-3.8	1.7-3.9	1.7-3.9	1.6-4.6	1.5-4.6	1.3-5.8	1.3-5.7		
	J	3.3	3.2	3.3	3.2	3.4	3.4	3.9	4.0	5.3	5.3
2.4	0.5	3.2-3.3	3.3-3.4	3.2-3.3	3.3-3.4	3.4-3.5	3.4-3.5	3.9-4.1	3.9-4.0	5.0-5.6	4.9-5.6
	1.0	3.4-3.6	3.4-3.6	3.4-3.6	3.5-3.6	3.4-3.8	3.4-3.9	3.7-4.4	3.6-4.8	3.6-7.9	3.6-7.9
	1.5	3.2-3.7	3.2-3.8	2.8-4.0	2.8-4.1	2.2-4.7	2.3-4.8	2.2-6.0	2.2-6.1	1.7-9.0	1.7-8.9
	J	4.5	4.5	4.5	4.5	4.5	4.5	4.8	4.8	5.8	5.9
2.8	0.5	4.4-4.5	4.5-4.6	4.4-4.5	4.5-4.6	4.4-4.5	4.5-4.6	4.7-4.9	4.8-4.9	5.5-5.8	5.6-5.7
	1.0	4.5-4.6	4.5-4.7	4.5-4.6	4.6-4.7	4.5-4.7	4.6-4.8	4.5-5.1	4.5-5.3	5.1-6.2	5.1-6.4
	1.5	3.8-4.7	3.9-4.7	3.9-4.7	4.0-4.7	3.6-5.0	3.7-5.0	3.2-5.7	3.2-5.8	3.3-7.5	3.3-7.6
	J	5.6	5.6	5.6	5.6	5.5	5.5	5.7	5.7	6.3	6.4
3.2	0.5	5.4-5.5	5.6-5.7	5.4-5.5	5.5-5.6	5.3-5.5	5.4-5.6	5.5-5.6	5.6-5.7	6.1-6.2	6.2-6.3
	1.0	5.4-5.5	5.5-5.6	5.4-5.5	5.5-5.6	5.4-5.5	5.5-5.6	5.4-5.7	5.5-5.8	5.7-6.4	5.6-6.6
	1.5	4.3-5.7	4.4-5.7	4.3-5.7	4.5-5.7	4.3-5.6	4.5-5.7	4.0-6.1	4.2-6.2	4.1-7.2	4.1-7.3
	J	6.5	6.5	6.4	6.5	6.4	6.4	6.6	6.6	6.9	6.9
3.6	0.5	6.4-6.5	6.4-6.5	6.4-6.5	6.4-6.4	6.3-6.4	6.3-6.4	6.5-6.6	6.5-6.6	6.7-6.8	6.7-6.8
	1.0	6.2-6.4	6.3-6.4	6.2-6.4	6.2-6.3	6.1-6.4	6.1-6.3	6.3-6.5	6.3-6.5	6.4-6.8	6.3-6.8
	1.5	4.7-6.6	4.7-6.8	4.7-6.5	4.8-6.7	4.8-6.6	4.8-6.5	4.6-6.9	4.6-6.9	4.6-7.3	4.6-7.3
	J	7.9	7.9	7.9	7.9	7.9	7.8	7.8	7.7	7.8	7.8
4.4	0.5	7.7-7.8	7.8-7.9	7.6-7.8	7.8-7.9	7.6-7.7	7.7-7.8	7.5-7.6	7.6-7.8	7.5-7.6	7.6-7.8
	1.0	7.4-7.6	7.5-7.7	7.4-7.6	7.4-7.7	7.3-7.5	7.4-7.6	7.3-7.4	7.3-7.5	7.3-7.5	7.4-7.5
	1.5	5.4-8.1	5.4-8.1	5.4-8.0	5.4-8.0	5.4-7.9	5.6-8.0	5.4-7.9	5.6-7.9	5.4-8.0	5.6-8.0

Table C.11

Range of $\alpha_{JR}(\%)$ values found for 11 values of c in an F–test for equality of two variances for n = 25 and α = 0.05, where r = 1.5, 1.0 and 0.5. J gives the level of significance ($\alpha_J\%$) where the Johnson distribution is not subject to rounding.

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
	J	1.4	1.3	1.4	1.4	1.5	1.6	1.8	1.9		
2.0	0.5	1.4-1.5	1.4-1.6	1.4-1.6	1.4-1.6	1.6-1.7	1.6-1.7	1.6-2.2	1.7-2.3		
	1.0	1.2-2.3	1.3-2.3	1.2-2.4	1.2-2.4	1.2-2.8	1.3-2.7	1.1-3.7	1.2-3.8	-	
	1.5	1.1-4.0	1.1-4.0	1.1-4.2	1.1-4.2	0.9-5.1	0.9-5.3	0.6-7.5	0.6-7.6		
	J	3.0	3.0	3.0	3.0	3.1	3.1	3.3	3.3	3.5	3.6
2.4	0.5	3.0-3.1	2.8-2.9	3.0-3.1	3.0-3.1	3.0-3.1	3.0-3.1	3.3-3.5	3.3-3.5	3.4-3.7	3.5-3.8
	1.0	3.2-3.4	3.1-3.2	3.2-3.4	3.2-3.4	3.0-3.5	3.0-3.5	3.2-3.9	3.2-3.9	2.2-5.7	2.3-6.0
	1.5	2.7-4.0	2.7-4.1	2.5-4.2	2.5-4.3	2.1-4.8	2.2-4.9	1.7-6.0	1.8-6.0	0.9-8.6	0.9-8.7
	J	4.2	4.3	4.3	4.3	4.4	4.3	4.5	4.5	4.8	4.9
2.8	0.5	4.2-4.3	4.3-4.4	4.3-4.4	4.3-4.4	4.3-4.4	4.3-4.4	4.4-4.6	4.4-4.6	4.7-4.9	4.7-5.0
	1.0	4.3-4.5	4.4-4.6	4.3-4.4	4.4-4.5	4.3-4.5	4.4-4.5	4.3-4.8	4.3-4.9	4.2-5.3	4.4-5.6
	1.5	4.3-4.4	4.3-4.5	4.2-4.6	4.3-4.7	3.7-5.1	3.8-5.1	3.4-5.8	3.5-5.8	3.2-7.4	3.3-7.6
	J	5.7	5.7	5.6	5.7	5.6	5.7	5.7	5.8	6.1	6.1
3.2	0.5	5.6-5.7	5.6-5.7	5.6-5.7	5.6-5.8	5.5-5.7	5.6-5.7	5.7-5.8	5.7-5.8	6.0-6.1	6.0-6.2
	1.0	5.5-5.7	5.6-5.7	5.5-5.6	5.6-5.7	5.5-5.6	5.5-5.7	5.5-5.8	5.6-5.8	5.5-6.3	5.7-6.4
	1.5	5.1-5.7	5.1-5.8	5.1-5.7	5.2-5.7	5.0-5.8	5.1-5.8	4.5-6.4	4.5-6.4	4.3-7.4	4.4-7.5
	J	6.9	7.0	6.9	7.0	6.8	6.9	6.9	7.0	7.2	7.3
3.6	0.5	6.8-6.9	6.9-7.0	6.7-6.9	6.8-7.0	6.7-6.8	6.7-6.8	6.8-6.9	6.9-7.0	7.1-7.2	7.1-7.3
	1.0	6.6-6.7	6.7-6.8	6.4-6.7	6.6-6.8	6.5-6.7	6.5-6.7	6.4-6.9	6.4-6.9	6.7-7.0	6.7-7.2
	1.5	5.7-6.8	5.8-6.8	5.7-6.8	5.8-6.8	5.7-6.8	5.7-6.8	5.4-7.2	5.3-7.2	5.2-7.9	5.3-7.9
	J	9.0	9.1	9.0	9.1	8.9	9.0	8.9	9.0	8.9	9.0
4.4	0.5	8.7-8.8	8.8-8.9	8.7-8.8	8.8-8.9	8.6-8.7	8.8-8.9	8.6-8.7	8.7-8.9	8.7-8.8	8.7-8.9
	1.0	8.4-8.5	8.4-8.6	8.3-8.5	8.4-8.7	8.2-8.4	8.3-8.5	8.2-8.4	8.3-8.5	8.3-8.4	8.4-8.6
	1.5	7.0-8.5	7.0-8.7	7.0-8.5	7.1-8.7	6.9-8.5	7.0-8.6	6.9-8.5	7.0-8.6	6.8-8.7	6.9-8.7

Table C.12

Range of $P_{JR}(\%)$ values found for 11 values of c in an F-test for equality of two variances for n = 10 and P = 0.3 and 0.7, where r = 0.5 and 1.0. J gives the level of power ($P_J\%$) where the Johnson distribution is not subject to rounding.

P = 0.3

	$\sqrt{\beta_1}$	0.0		0.2		0.4		0.6		0.8	
β_2	r	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail	lower tail	upper tail
2.0	J	26.8	26.6	27.0	26.7	27.3	26.9	27.1	27.4	26.8	27.4
	0.5	25.5-26.3	25.8-26.5	25.9-26.4	25.7-26.3	25.5-26.9	26.1-26.7	24.7-27.5	25.6-27.8	19.0-32.6	25.7-29.6
	1.0	23.1-24.5	24.5-25.2	22.2-25.0	24.4-24.8	20.2-28.2	24.1-25.8	14.5-36.6	21.8-30.6	13.6-48.0	17.8-40.7
3.6	J	31.9	30.7	31.7	30.8	31.5	31.0	31.7	31.5	31.9	31.9
	0.5	30.3-31.1	29.8-30.4	30.2-31.1	29.8-30.4	29.1-30.8	30.2-30.5	29.6-30.8	30.3-30.4	30.3-30.9	30.9-31.3
	1.0	26.4-28.0	28.2-28.9	26.8-28.0	28.1-29.0	26.5-27.5	28.2-29.0	25.3-28.2	28.3-29.5	24.1-30.7	28.5-30.1

P = 0.7

2.4	J	73.4	73.2	73.3	73.0	73.1	72.7	72.6	72.9	72.6	73.2
	0.5	69.1-70.0	71.9-72.1	68.6-69.6	71.8-72.3	67.4-70.3	71.7-72.3	64.6-73.7	71.7-72.8	64.2-78.6	69.6-72.2
	1.0	54.5-63.5	68.9-69.3	52.4-65.8	68.6-69.8	48.8-72.1	68.1-70.2	38.8-82.7	65.6-73.2	29.9-87.4	65.2-79.1
3.6	J	69.3	68.1	69.2	68.3	69.1	68.5	68.5	68.3	68.1	68.1
	0.5	65.5-66.2	67.1-67.6	65.4-65.9	67.2-68.8	65.4-66.0	67.3-68.0	65.3-65.9	67.2-67.8	64.2-65.8	67.1-67.6
	1.0	52.0-60.1	64.9-65.4	52.0-60.0	65.1-65.6	51.7-59.5	65.3-65.7	49.1-63.2	64.9-65.6	47.2-66.5	64.5-65.7

Table C.13

Range of $P_{JR}(\%)$ values found for 11 values of c, in an F-test for equality of two variances for n = 25 and P = 0.3 and 0.7 (lower tail), where r = 0.5, 1.0 and 1.5. J gives the level of power ($P_J\%$) where the Johnson distribution is not subject to rounding.

P = 0.3

β_2	$\sqrt{\beta_1}$	0.0	0.2	0.4	0.6	0.8
		lower tail	lower tail	lower tail	lower tail	lower tail
2.4	J	27.4	27.2	27.0	26.8	26.4
	0.5	26.0-26.5	25.9-26.8	25.5-26.7	23.6-29.3	17.6-37.2
	1.0	23.3-24.5	22.6-25.7	19.7-28.5	14.2-41.1	11.7-62.3
	1.5	16.9-25.2	14.4-28.3	10.1-37.2	7.4-56.9	4.6-89.7
3.6	J	32.0	32.0	31.8	32.0	32.3
	0.5	30.7-31.7	30.5-31.7	30.4-31.1	30.7-31.3	31.0-31.4
	1.0	28.1-28.8	27.8-29.2	27.7-28.8	27.0-29.0	26.1-31.0
	1.5	20.8-28.0	20.6-28.1	20.6-28.3	18.1-33.1	15.4-35.6

P = 0.7

2.4	J	73.9	73.6	73.4	73.4	73.1
	0.5	70.3-70.9	70.1-70.8	69.2-72.3	68.0-73.4	58.2-78.9
	1.0	61.6-63.3	59.9-64.5	54.8-70.8	43.1-82.9	35.9-94.0
	1.5	50.1-52.7	41.7-63.7	29.6-77.7	25.1-91.9	20.0-97.8
3.6	J	68.7	68.7	68.5	68.4	68.4
	0.5	65.9-66.9	65.7-66.7	65.9-66.6	65.7-66.4	65.6-66.3
	1.0	59.4-60.6	59.1-60.6	58.8-60.9	56.1-61.3	54.6-65.6
	1.5	37.2-62.1	37.2-61.9	37.3-62.8	31.4-67.3	27.4-72.9

Table C.14

Range of $\alpha_{JR}(\%)$ values found for 11 values of c in a one-way analysis of variance for k = 3 and n = 10, where α = 0.05. J gives the level of significance ($\alpha_J\%$) where the Johnson distribution is not subject to rounding.

β_2	r \ $\sqrt{\beta_1}$	0.0	0.4	0.8
2.0	J	5.1	5.1	
	1.0	5.0 - 5.2	4.9 - 5.1	
	1.5	4.8 - 5.3	4.7 - 5.3	———
	2.0	4.2 - 5.4	4.2 - 5.4	
2.4	J	5.0	5.0	4.9
	1.0	4.9 - 5.1	4.9 - 5.1	4.8 - 5.1
	1.5	4.8 - 5.2	4.7 - 5.2	4.4 - 5.2
	2.0	4.7 - 5.4	4.5 - 5.4	4.5 - 5.3
4.4	J	4.7	4.8	4.7
	1.0	4.7 - 4.8	4.7 - 4.9	4.7 - 4.8
	1.5	4.7 - 5.1	4.6 - 5.1	4.6 - 5.0
	2.0	4.7 - 5.2	4.7 - 5.2	4.5 - 5.2

Table C.15

Range of the significance level ($\alpha_R\%$) found for 11 values of c in a one way analysis of variance for k = 3 and n = 10, where α = 0.05 for rounded normal data.

r	$\alpha_R(\%)$
1.0	4.7 - 5.1
1.5	4.5 - 5.1
2.0	4.4 - 5.1

Table C16

Range of $P_{JR}(\%)$ values found for 11 values of c in a one way analysis of variance for k = 3 and n = 10 and P = 0.3 and 0.7, where r = 0.5, 1.0, 1.5 and 2.0. J gives the level of power ($P_J\%$) where the Johnson distribution is not subject to rounding.

	$\sqrt{\beta_1}$	0.0		0.4		0.8	
β_2	r	P=0.3	P=0.7	P=0.3	P=0.7	P=0.3	P=0.7
	J	29.1	69.7	29.2	69.9		
2.0	0.5	28.3-28.9	68.5-69.1	28.2-29.6	68.5-69.5		
	1.0	26.5-28.2	64.1-67.5	25.2-30.1	62.2-70.3	———	
	1.5	23.0-28.8	60.0-64.7	19.4-31.4	57.6-68.9		
	2.0	17.9-28.0	48.2-66.1	13.9-34.7	41.9-77.1		
	J	29.4	70.3	29.5	70.0	29.8	69.8
2.4	0.5	28.7-29.3	68.8-69.3	28.6-29.2	68.7-69.4	27.1-32.6	66.3-71.6
	1.0	27.2-27.7	66.0-66.3	27.2-28.2	65.4-67.0	21.1-35.8	59.7-75.8
	1.5	25.1-26.1	61.2-62.5	24.5-27.1	60.2-64.5	12.4-45.7	48.9-79.4
	2.0	21.8-24.5	54.2-59.8	18.7-27.2	50.4-63.7	10.7-49.0	28.3-83.0
	J	30.9	71.4	30.9	71.4	31.1	71.8
4.4	0.5	30.3-30.8	70.3-70.9	30.2-30.8	70.3-70.9	30.3-30.7	70.4-71.1
	1.0	28.3-29.5	67.4-68.3	28.2-29.4	67.5-68.5	28.4-30.6	67.5-71.2
	1.5	26.0-26.9	62.8-63.8	25.9-26.9	62.6-63.8	25.6-27.7	62.3-64.8
	2.0	21.5-25.8	54.6-60.6	21.4-25.9	54.5-60.8	21.6-27.1	54.1-62.8

Table C.17

Range in power (P_R) found for 11 values of c in a one way analysis of variance for k = 3 and n = 10, for P = 0.3 and 0.7 where α = 0.05, for rounded normal data.

r	P = 0.3	P = 0.7
0.5	28.7 - 29.3	68.8 - 69.3
1.0	27.2 - 27.7	66.0 - 66.3
1.5	25.4 - 26.5	61.7 - 62.8
2.0	22.8 - 24.1	56.4 - 57.8

Approximation Formulae for the mean and variance of functions of random variables X and Y

In general, there are no simple exact formulae for the mean and variance of functions of random variables X and Y; however there are approximate formulae which are sometimes useful. We have from Mood, Graybill and Boes (1974) approximate formulae for the mean and variance of a function $g(X_1,X_2)$

$$E[g(X_1,X_2)] \approx g(\mu_1,\mu_2) + \tfrac{1}{2}\sigma_1^2\left[\frac{\partial^2 g}{\partial x_1^2}\right]_{\substack{x_1=\mu_1\\x_2=\mu_2}} + \tfrac{1}{2}\sigma_2^2\left[\frac{\partial^2 g}{\partial x_2^2}\right]_{\substack{x_1=\mu_1\\x_2=\mu_2}} + \mathrm{Cov}(X_1,X_2)\left[\frac{\partial^2 g}{\partial x_1 \partial x_2}\right]_{\substack{x_1=\mu_1\\x_2=\mu_2}} \qquad (1)$$

$$V[g(X_1,X_2)] \approx \sigma_1^2\left[\left[\frac{\partial g}{\partial x_1}\right]_{\substack{x_1=\mu_1\\x_2=\mu_2}}\right]^2 + \sigma_2^2\left[\left[\frac{\partial g}{\partial x_2}\right]_{\substack{x_1=\mu_1\\x_2=\mu_2}}\right]^2 + 2\mathrm{Cov}(X_1,X_2)\left[\frac{\partial g}{\partial x_1}\right]_{\substack{x_1=\mu_1\\x_2=\mu_2}} \times \left[\frac{\partial g}{\partial x_2}\right]_{\substack{x_1=\mu_1\\x_2=\mu_2}} \qquad (2)$$

where the mean and variance of X_1 and X_2 are respectively (μ_1,σ_1^2) and (μ_2,σ_2^2).

Using (1) and (2) above we can obtain approximate expressions of the mean and variance of t_R (5.3-2). From (5.3-2) we have

$$t_R = \frac{\bar{X}_R}{S_R/\sqrt{n}}$$

Consider the function

$$g(X_1,X_2) = \frac{X_1}{\sqrt{X_2}} \qquad \text{where } X_1 = \bar{X}_R,\ X_2 = S^2_R$$

We have

$$\mu_1 = E[X_1] = E[\bar{X}_R] = \mu_R \qquad \mu_2 = E[X_2] = E[S_R^2] = \sigma_R^2$$

$$\sigma_1^2 = V[X_1] = V[\bar{X}_R] = \frac{\sigma_R^2}{n} \qquad \sigma_2^2 = V[X_2] = V[S_R^2]$$

$$= \left[\frac{2}{n-1} + \left[\frac{\beta_{2R}-3}{n}\right]\right]\sigma_R^4$$

From Kendall and Stuart (1968, pp233)

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(\bar{X}_R, S_R^2) = \frac{\mu_{3R}}{n} \qquad \text{to order } n^{-1}.$$

Using (1) we have

$$E[g(X_1, X_2)] = E\left[\frac{\bar{X}_R}{S_R^2}\right] \approx \frac{\mu_R}{\sigma_R}\left[1 + \frac{3}{8}(\beta_{2R}-1)/n\right] - \frac{1}{2n}\sqrt{\beta_{1R}}$$

$$V[g(X_1, X_2)] = V\left[\frac{\bar{X}_R}{S_R^2}\right] \approx \frac{1}{n} + \frac{1}{4}\frac{\mu_R^2}{\sigma_R^2}\left[(\beta_{2R}-1)/n\right] - \frac{\mu_R}{\sigma_R}\frac{\sqrt{\beta_{1R}}}{n}$$

hence $$E[t_R] = \sqrt{n}\, E\left[\frac{\bar{X}_R}{S_R^2}\right] \approx \frac{\mu_R\sqrt{n}}{\sigma_R}\left[1 + \frac{3}{8}(\beta_{2R}-1)/n\right] - \frac{1}{2n}\sqrt{\beta_{1R}}$$

$$V[t_R] = nV\left[\frac{\bar{X}_R}{S_R^2}\right] \approx 1 + \frac{1}{4}\frac{\mu_R^2}{\sigma_R^2}(\beta_{2R}-1) - \frac{\mu_R}{\sigma_R}\sqrt{\beta_{1R}}$$

ABERNETHY J (1933): On the elimination of systematic errors due to grouping: Annals of Mathematical Statistics 4, 263–277.

AIGNER D and GOLDBERGER A (1970): Estimation of Pareto's law from grouped observations: Journal of the American Statistical Association 65, 712–723.

ANON (1975): Guide to Statistical Interpretation of Data Part 1. Routine Analysis of Quantitative Data, BS 2846: British Standards Institution, 2 Park Street, London.

BATEN WD (1931): Correction for the moments of a frequency distribution in two variables: Annals of Mathematical Statistics 2, 309–319.

BEATON AE, RUBIN DB and BARONE JL (1976): The acceptability of regression solutions : another look at computational accuracy: Journal of the Americal Statistical Association 71, 158–168.

BENN R and SIDEBOTTOM S (1976): Algorithm AS95. Maximum likelihood estimation of location and scale parameters from grouped data: Applied Statistics 28, 88–93.

BERNARDO JM (1976): Algorithm AS103. Psi (digamma) function: Applied Statistics 25, 315–317.

BOX GEP (1954): Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one–way classification: Annals of Mathematical Statistics 25, 290–302.

BURRIDGE J (1981): A note on maximum likelihood estimation for regression models using grouped data: Journal of the Royal Statistical Society B43, 41–45.

BURRIDGE J (1982): Some unimodality properties of likelihoods derived from grouped data: Biometrika 69, 145–151.

BURRIDGE J (1986): Existence of maximum likelihood estimates in regression models for grouped and ungrouped data: Journal of the Royal Statistical Society B48, 100–106.

CAMERON TA (1987): The impact of grouping coarseness in alternative grouped–date regression models: Journal of Econometrics 35, 37–57.

CHURCH AER (1925): On the moments of the distribution of squared standard deviations for samples of N drawn from an indefinitely large population: Biometrika 17, 79–83.

COCHRAN WG and COX GM (1957): Experimental Design, 2nd edn, Wiley, New York.

CORNISH E and FISHER R (1937): Moments and cumulants in the specification of distributions: Review of the International Statistical Institute 5, 307–320.

CRAIG C (1941): A note on Sheppard's correction: Annals of Mathematical Statistics 12, 339–345.

DAVID FN and JOHNSON NL (1951): The effect of non-normality on the power function of the F-test in the analysis of variance: Biometrika 38, 43-57.

DAVIES G and BRUNER N (1943): A second moment correction for grouping: Journal of the American Statistical Association 38, 63-68.

DEKEN J (1983): Approximating conditional moments of the multivariate normal distribution: SIAM Journal of Scientific and Statistical Computing 4, 720-732.

DEMPSTER AP, LAIRD N and RUBIN DB (1977): Maximum likelihood from incomplete data via the EM algorithm: Journal of the Royal Statistical Society B39, 1-38.

DEMPSTER AP and RUBIN DB (1983): Rounding error in regression : the appropriateness of Sheppard's corrections: Journal of the Royal Statistical Society B45, 51-59.

DON FJ (1981): A note on Sheppard's correction for grouping and maximum likelihood estimation: Journal of Multivariate Analysis 11, 452-458.

DURBIN J (1954): Errors in variables: Review of the International Statistical Institute 22, 23-32.

DYKE GV (1974): Comparative Experiments with Field Crops: London, Butterworths.

DYM H and McKEAN H (1972): Fourier Series and Integrals: Academic Press.

EISENHART C (1947): Effects of rounding or grouping data: Chapter 4 of Techniques of Statistical Analysis, ed C Eisenhart, MW Hastay and WA Wallis: McGraw-Hill, New York.

FISHER RA (1922): On the mathematical foundations of theoretical statistics: Philosophical Transactions of the Royal Society A222, 309-368.

FISHER RA (1925): Theory of statistical estimation: Proceedings of the Cambridge Philosophical Society 22, 700-710

FISHER RA (1936): Statistical Methods for Research Workers, 6th ed: Oliver and Boyd, Edinburgh.

FRYER JG and PETHYBRIDGE RJ (1972): Maximum likelihood estimation of a linear regression function with grouped data: Applied Statistics 21, 142-154.

GAYEN AK (1950): The distribution of the variance ratio in random samples of any size drawn from non-normal universes: Biometrika 37, 236-255.

GEARY RC (1947): Testing for normality: Biometrika 34, 209-242.

GHURYE SG (1949): On the use of student's t-test in an asymmetrical population: Biometrika 36, 426-430.

GJEDDEBAEK NF (1949): Contributions to the study of grouped observations. Application of the method of maximum likelihood in case of normally distributed obversations: Skandinavisk Aktuarietidskrift 32, 135–159.

GJEDDEBAEK NF (1956): Contributions to the study of grouped observations. II. Loss of information caused by grouping of normally distributed observations: Skandinavisk Aktuarietidskrift 39, 154–159.

GJEDDEBAEK NF (1957): Contributions to the study of grouped observations. III. The distribution of estimates of the mean: Skandinavisk Aktuarietidskrift 40, 20–25.

GJEDDEBAEK NF (1959): Contributions to the study of grouped observations. IV. Some comments on simple estimates: Biometrics 15, 433–439.

GJEDDEBAEK NF (1968): Statistical analysis : grouped observatons: International Encyclopaedia of the Social Sciences 15, ed DR Sills: Macmillan and the Free Press, New York.

GRUNDY P (1952): The fitting of grouped truncated and grouped censored normal distributions: Biometrika 39, 252–259.

HAITOVSKY Y (1973): Regression Estimation from Grouped Observations: Griffin, London.

HARTLEY HO (1950): A simplified form of Sheppard's correction formulae: Biometrika 37, 145–148.

HOLLAND B (1975): Some results on the discretization of continuous probability distributions: Technometrices 17, 333–339.

HUGHES HM (1949): Estimation of the variance of the bivariate normal distribution: University California Publications Statistics 1, 37–52.

INDRAYAN A and RUSTAGI JS (1979): Approximate maximum likelihood estimates in regression models for grouped data: Optimizing Methods in Statistics, editor JS Rustagi: Academic Press, New York, pp301–319.

JOHNSON NL (1949): Systems of frequency curves generated by methods of translation: Biometrika 36, 149–176.

JOHNSON NL and KOTZ S (1970): Continuous Univariate Distributions, Vol 2: Wiley, New York.

KANJI GK (1976): Effect of non–normality on the power in analysis of variance : a simulation study: International Journal of Mathematical Education and Science Technology 7, 155–160.

KANJI GK (1977): Power aspects of analysis of variance in random effects models : a simulation study: International Journal of Mathematical Education and Science Technology 8, 293–297.

KAWATA T (1940): On the division of probability laws: Proceedings Imperial Academy Tokyo 16, 249–254.

KENDALL MG (1938): The conditions under which Sheppard's Corrections are valid: Journal of the Royal Statistical Society 101, 592–605.

KENDALL MG and STUART A (1968): The Advanced Theory of Statistics, Vol 1: Griffin, London.

KRUSKAL W and TANUR J (1978): Statistical analysis : grouped observations: International Encyclopaedia of Statistics: Macmillan Free Press, New York; Collier Macmillan, London.

KRUTCHOFF RG (1967): Letter to the editor: American Statistician 21, 35.

KULLBACK S (1935): A note on Sheppard's corrections: Annals of Mathematical Statistics 6, 158–159.

KULLDORFF G (1961): Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples: Stockholm : Almqvist and Wiksell, New York : John Wiley and Sons Inc.

LANGDON WH and ORE O (1930): Semi–invariants and Sheppard's corrections: Annals of Mathematics 31, 230–232.

LAWLESS JF (1982): Statistical Models and Methods for Lifetime Data: New York : Wiley.

LAU C (1980): Algorithm AS147. A simple series for the incomplete gamma integral: Applied Statistics 29, 113–114.

LEWIS WT (1935): A reconsideration of Sheppard's corrections: Annals of Mathematical Statistics 6, 11–20.

LINDLEY DV (1950): Grouping corrections and maximum likelihood equations: Proceedings of the Cambridge Philosophical Society 46, 106–110.

LONGLEY JW (1967): An appraisal of least squares programs for the electronic computer from the point of view of the user: Journal of the American Statistical Association 62, 819–841.

LOWELL M (1980): On round–off error: Analytical Chemistry 52, 1141–1147.

LUKAS E (1960): Characteristic Functions, 2nd ed: Griffin, London.

MARTIN E (1934): On the corrections for the moment coefficients of frequency distributions when the start of the frequency is one of the characteristics to be determined: Biometrika 26, 12–58.

McKAY AT (1932): A Bessel function distribution: Biometrika 24, 39–44.

McNEIL DR (1966): Consistent statistics for estimating and testing hypothesis from grouped samples: Biometrika 53, 545–557.

MOOD AM, GRAYBILL FA and BOES DC (1974): Introduction to the Theory of Statistics, 3rd edn: McGraw-Hill, New York.

NICHOLSON MD (1979): On expressing the mean of rounded data: Biometrics 35, 873-874.

NORTON V (1983): A simple algorithm for computing the non-central F-distribution: Applied Statistics 32, 84-85.

OWEN DB (1965): The power of Student's t-test: Journal of the American Statistical Association 60, 320-333.

PAIRMAN E and PEARSON K (1918): On correcting for the moment - coefficients of limited range frequency - distributions when there are finite or infinite ordinates and any slopes at the terminals of the range: Biometrika 12, 231-258.

PEARSE G (1928): On the corrections for the moment coefficients of frequency distributions when there are infinite ordinates at one or both terminals of the range: Biometrika 20A, 314-355.

PEARSON ES (1931): Analysis of variance in case of non-normal variation: Biometrika 23, 114-133.

PEARSON ES and HARTLEY HO (1972): Biometrika Tables for Statisticians, Vol 2: Cambridge University Press.

PEARSON ES and PLEASE NW (1975): Relation between the shape of population distribution and the robustness of four test statistics: Biometrika 62, 223-240.

PEARSON K (1902): On the systematic fitting of curves to observations and measurements: Biometrika 1, 265-303.

PETHYBRIDGE RJ (1973): Maximum likelihood estimation of a polynomial regression function with grouped data: Applied Statistics 22, 203-212.

PETHYBRIDGE RJ (1975): Maximum likelihood estimation of a linear regression function with grouped data: Applied Statistics 24, 28-41.

POSTEN HO (1978): The robustness of the two sample t-test over the Pearson system: Journal of Statistical Computation and Simulation 6, 295-311.

POSTEN HO (1979): The robustness of the one sample t-test over the Pearson system: Journal of Statistical Computation and Simulation 9, 133-149.

PREECE DA (1982): t is for trouble (and textbooks) : A critique of some examples of the paired-sample t-test: The Statistician 31, 169-195.

RILEY J, BEKELE I and SHREWSBURY B (1983): How an analysis of variance is affected by the degree of precision of the data: BIAS 10, 18-42.

SANDON F (1924): None on the simplification of the calculation of abruptness coefficients to correct crude moments: Biometrika 16, 193–195.

SCHADER M and SCHMID F (1984): Computation of maximum likelihood estimates of μ and σ from grouped sample of a normal population. A comparison of algorithms: Statische Hefte 25, 245–258.

SHEPPARD WF (1898): On the calculation of the most probable values of frequency constraints for data arranged according to equidistant divisions of scale: Proceedings of the London Mathematical Society 29, 353–380.

SNEDECOR GW and COCHRAN WG (1967): Statistical Methods, 6th edn: Iowa State University Press, Ames.

SRIVASTAVA ABI (1958): Effect of non–normality on the power function of t–test: Biometrika 45, 421–429.

STUDENT (1908): The probable error of a mean: Biometrika 6, 1–25.

SWAMY P (1960): Estimating the mean and variance of a normal distribution from singly and doubly truncated samples of grouped observations: Calcutta Statistical Association Bulletin 9, 145–156.

SWAN A (1969): Algorithm AS16. Maximum likelihood estimation from grouped and censored normal data: Applied Statistics 18, 110–114.

SWINDEL FB and BOWER DR (1972): Rounding errors in the independent variables in a general linear model: Technometrics 14, 215–218.

TALLIS G and YOUNG S (1962): Maximum likelihood estimation of parameters of the normal, lognormal, truncated normal and bivariate normal distributions from grouped data: The Australian Journal of Statistics 4, 49–54.

TALLIS G (1967): Approximate maximum likelihood from grouped data: Technometrics 9, 599–606.

TIKU ML (1967): Tables of the power of the F–test: Journal of the American Statistical Association 62, 525–539.

TOCHER KD (1949): A note on the analysis of grouped probit data: Biometrika 36, 9–17.

TRICKER AR (1984a): Effects of rounding data sampled from the exponential distribution: Journal of Applied Statistics 11(1), 51–87.

TRICKER AR (1984b): Effects of rounding on the moments of a probability distribution: The Statistician 33, 381–390.

VAN–WAERDEN B (1973): Mathematische Statistik, 3rd edition: Springer Berlin.

WATTS D (1961): A general theory of amplitude quantization with applications to correlation determination: Proceedings of the Institute of Electrical Engineers ptC 109, 209–218.

WIDROW B (1956): A study of rough amplitude quantization by means of nyquist sampling theory: Transactions of the Institute of Radio Engineers CT-3, 266-276.

WIDROW B (1961): Statistical analysis of amplitude quantiz sampled data systems: Transactions of the American Institute of Electrical Engineers 77 pt2, 555-568.

WOLD H (1934): Sheppard's correction formulae in several variables: Skandinavisk Aktuarietidskrift 17, 248-255.

WOLYNETZ M (1979a): Algorithm AS138. Maximum likelihood estimation from confined and censored normal data: Applied Statistics 28, 185-195.

WOLYNETZ M (1979b): Algorithm AS139. Maximum likelihood estimation in a linear model from confined and censored normal data: Applied Statistics 28, 195-206.

YATES F (1937): The design and analysis of factorial experiments: Technical Communication No 35 of the Commonwealth Bureau of Soils, Harpenden, England.

YONEDA K and UCHIYAMA M (1956): Some estimations in the case of relatively large class intervals: Yokohama Mathematical Journal 4, 99-118.

The Statistician 33 (1984) 381–390

Effects of Rounding on the Moments of a Probability Distribution

A. R. TRICKER

Department of Mathematics, Statistics and Operational Research, Sheffield City Polytechnic

Abstract: This paper looks at the effects of rounding data sampled from a probability distribution. Using the characteristic function of the rounded observations, the influence of the rounding process on the first two moments is examined. The normal, Laplace and gamma distributions are considered. The results indicate that both the degree of rounding and the skewness of a distribution are important in determining how much the mean and variance are distorted by the rounding process.

1 Introduction

We are often faced with the problem that data sampled from continuous distributions have been rounded. The reduction in precision of the data can sometimes distort the information conveyed by the measurements. Some discussion of this topic has appeared in Eisenhart (1947), Fisher (1922), Gjedderbaek (1968) and Lowell (1980). Most of these works are confined to the normal distribution. Only Lowell (1980) deals specifically with the effect of rounded data on the moments of the normal distribution. In the present paper the necessary theory is developed to allow the moments of any distribution to be investigated. By deriving the characteristics function of the rounded distribution, general expressions for its mean and variance are obtained. This makes it easy to measure the amount of distortion caused by the rounding process on the mean and variance. Both symmetrical and non-symmetrical distributions are considered. To the author's knowledge nothing has been written about the association between skewness and the rounding process. The change in moments caused by rounding, together with the effect of skewness, is shown for the normal, Laplace and gamma distributions.

2 Characteristic function of the rounded distribution

If values from a continuous random variable X are rounded, the result is a new discrete random variable X'. Let x and x' represent values of the random variables X and X' respectively. If values of x are rounded into intervals of width ω, with midpoints x', and the centre of the interval containing zero is $a\omega$, then x' has the following values.

$$a\omega, a\omega \pm \omega, a\omega \pm 2\omega, \ldots \tag{2.1}$$

(2.1) will be known as the rounding lattice. Here a determines the position of the rounding lattice and may be located at random between $-\omega/2$ and $\omega/2$. Thus the mathematical

relationship between x and x' is such that if:

$$a\omega+(n-\tfrac{1}{2})\omega<x'\leq a\omega+(n+\tfrac{1}{2})\omega$$

then

$$x'=(a+n)\omega, \qquad n=0,\pm1,\pm2,\ldots \tag{2.2}$$

The relationship (2.2) can easily be adapted for where the random variable is defined only for positive values of x. Watts (1961) derived the characteristic function of the general quantizer system for an electric signal. By letting the gain and shift be equal to ω and a respectively in the quantizer system, we can obtain the characteristic function $\phi_{X'}(t)$ of X given by:

$$\phi_{X'}(t)=\sum_{k=-\infty}^{\infty}\exp(1-i2\pi ka)\,\phi_X\left(t+\frac{2\pi k}{\omega}\right)\frac{\sin\tfrac{1}{2}(t\omega+2\pi k)}{\tfrac{1}{2}(t\omega+2\pi k)} \tag{2.3}$$

where $\phi_X(.)$ is the characteristic function of the continuous distribution X.

$$E[X'^s]=(-i)^s\left\{\frac{d^s\phi_{X'}(t)}{dt^s}\right\}_{t=0} \tag{2.4}$$

To derive the effect of rounding on the mean and variance, we require the first two moments of X'. If the operation indicated in (2.4) with $s=1$ and 2 is applied to (2.3), the first two moments of X' can be found after lengthy manipulations. General expressions will be obtained for the first two moments for a continuous distribution, which may be either symmetrical or non-symmetrical.

The results that follow are only valid for continuous distributions which are uniquely determined by their moments in which case (2.3) is a convergent series (e.g. Kendall & Stuart 1968).

Case 1: Continuous distribution – symmetrical about zero

If values from a continuous distribution symmetrical about zero are rounded into intervals of width ω, and the centre of the interval containing zero is $a\omega$, then we have the following results:

$$E(X')=-\frac{\omega}{\pi}\sum_{k=1}^{\infty}-\frac{(-1)^k}{k}\phi_X\frac{(2\pi k)}{\omega}\sin(2\pi ka) \tag{2.5}$$

$$E(X'^2)=E(X^2)+\frac{\omega^2}{12}+\sum_{k=1}^{\infty}(-1)^k\cos(2\pi ka)$$
$$X\left\{\left(\frac{\omega}{\pi k}\right)^2\phi_X\left(\frac{2\pi k}{\omega}\right)-\frac{2\omega}{\pi k}\phi'_X\left(\frac{2\pi k}{\omega}\right)\right\} \tag{2.6}$$

where $\phi_X(.)$ is the characteristic function of continuous random variable X

$$\phi'_X\left(\frac{2\pi k}{\omega}\right)=\left\{\frac{d}{dt}\phi_X(t)\right\}_{t=2\pi k/\omega}$$

Case 2: Continuous distribution – non-symmetrical

For a non-symmetrical distribution we have the following results.

$$E(X')=E(X)+\frac{\omega}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\{B\cos(2\pi ka)-A\sin(2\pi ka)\} \tag{2.7}$$

$$E(X'^2)=E(X^2)+\frac{\omega^2}{12}+\sum_{k=1}^{\infty}(-1)^k\left[\left(\frac{\omega}{\pi k}\right)^2\{A\cos(2\pi ka)+B\sin(2\pi ka)\}\right.$$
$$\left.-\frac{2\omega}{\pi k}\{A'\cos(2\pi ka)+B'\sin(2\pi ka)\}\right] \tag{2.8}$$

where

$\phi_X(2\pi k/\omega)=A+iB$

$\phi_{X'}(2\pi k/\omega)=A'+iB'$

When the condition (2.9) is placed on the characteristic function of X, only the central section of (2.3) enters in the calculation of the moments of X'.

$$\phi_X(t)=0 \quad |t|\geqslant\frac{2\pi}{\omega}$$

or

$$\phi_X(2\pi k/\omega)=0, \qquad k=\pm1,\pm2\ldots \tag{2.9}$$

The expression for the central section of (2.3) is

$$\phi(t)_{X'}\,|_{\text{central section}}=\phi_X(t)\frac{\sin(tw/2)}{tw/2} \tag{2.10}$$

The central section (2.10) can be thought of as a characteristic function in its own right. It is the product of the characteristic function of X and a variable which is uniformly distributed between $-\omega/2$ and $\omega/2$. Satisfaction of condition (2.9) suggests:

the moments of X' are the same as those of the sum of the moments of X and a statistically independent error, uniformly distributed on $(-\omega/2, \omega/2)$ (2.11)

The implications of (2.11) are: $E(X')=E(X)$; $V(X')=V(X)+\omega^2/12$.

The most common assumptions (Fisher, 1922; Eisenhart, 1947) concerning properties of rounding are those of (2.11). However their validity is often in doubt, as satisfaction of (2.9) is uncommon, the reason being that it is rare to have a probability distribution, whose characteristic function is zero, outside a finite range of t. The value of a characteristic function outside the region given by (2.9) is often very small and may be regarded as negligible for the accuracy we are interested in. As a result, (2.11) may be assumed to be true and the effect of the rounding process on the moments slight. In the following section we shall investigate whether this distortion in the moments may be considered negligible for certain distributions.

3 Symmetrical distributions – normal and Laplace

We consider the normal distribution first, to demonstrate how the moments may be distorted by rounding. Using the characteristic function of a normal distribution with mean zero and variance σ^2 we obtain the following from equations (2.5) and (2.6).

$$E(X')=-\frac{\omega}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}D\sin(2\pi ka) \tag{3.1}$$

$$E(X'^2)=\sigma^2+\frac{\omega^2}{12}+4\sum_{k=1}^{\infty}(-1)^k D\left\{\sigma^2+\left(\frac{\omega}{2\pi k}\right)^2\right\}\cos(2\pi ka) \tag{3.2}$$

where $D=\exp(-2k^2\pi^2\sigma^2/\omega^2)$ and $-\frac{1}{2}\leqslant a\leqslant\frac{1}{2}$.

If we let $\omega=r\sigma$, then r measures the degree of rounding with respect to the standard deviation σ. As r indicates the severity of rounding we express equations (3.1) and (3.2) in the form of r.

$$E(X')=-\sigma\left\{\frac{r}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k'}D'\sin(2\pi ka)\right\} \tag{3.3}$$

$$E(X'^2)=\sigma^2\left\{1+\frac{r^2}{12}\right\}+4\sum_{k=1}^{\infty}(-1)^k D'\left\{1+\left(\frac{r}{2\pi k}\right)^2\right\}\cos(2\pi ka) \tag{3.4}$$

where $D'=\exp(-2k^2\pi^2/r^2)$.

Of particular interest is the bias in $E(X')$. Equation (3.5) shows this bias relative to ω. The effect of rounding on the variance is best shown by expressing the $V(X')$ relative to $V(X)$ (equation (3.6)).

$$B=\frac{E(X')-E(X)}{\omega} \tag{3.5}$$

$$V=V(X')/V(X) \tag{3.6}$$

Figures 1 and 3 show curves for B and V for a ranging between $-\frac{1}{2}$ and $\frac{1}{2}$, and r up to 5.0.

The procedure for obtaining the first two moments for a normal distribution can also be used on the Laplace distribution, which has the following probability density.

$$f(x)=\frac{1}{2\beta}\exp\left(\frac{-|x|}{\beta}\right), \qquad -\infty\leqslant x\leqslant\infty \tag{3.7}$$

with mean zero and variance $\sigma^2=2\beta^2$.

Using the characteristic function of (3.7) we can obtain from (2.5) and (2.6) expressions for the first two moments in terms of r.

$$E(X')=-\sigma\left\{\frac{r}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}L\sin(2\pi ka)\right\} \tag{3.8}$$

$$E(X'^2)=\sigma^2\left[1+\frac{r^2}{12}+\sum_{k=1}^{\infty}(-1)^k L\left\{\left(\frac{r}{\pi k}\right)^2+4L\right\}\right] \tag{3.9}$$

$$=(1+2k^2\pi^2r^2)^{-1}$$

Figures 2 and 4 show curves for B and V respectively for the Laplace distribution, for a ranging between $-\frac{1}{2}$ and $\frac{1}{2}$ and r up to 5·0.

Figures 1–4 illustrate the effect of rounding on the two distributions. The graphs for B are symmetrical about $a=0$ and the bias is zero at $a=0$ and $a=\pm\frac{1}{2}$. In general, whenever the mean coincides with the boundary or centre of a rounding interval, then the bias is zero. Clearly the amount of bias caused by rounding is more severe in the Laplace distribution. For $r=1$ the maximum bias in the mean for the normal is $8{\cdot}1\ (10)^{-10}$ and for the Laplace is $1{\cdot}5\ (10)^{-2}$. For the variance (Figures 2 and 4) the effect of rounding is similar for a between 0·2 to 0·5 and −0·5 to −0·2. When a is between ±0·2 the $V(X')$ can be considerably lower than the $V(X)$. This is not true for the Laplace.

Of interest is the limiting behaviours of the expectation and variance of X' when the distribution is symmetrical. From equations (3·1–3·9) it can be shown that as r approaches zero, $E(X')$ and $V(X')$ tend to zero and σ^2 respectively. When r approaches ∞ the position of $a\omega$ is very important. Figure 5a shows when the distribution is between cell boundaries. All values will be rounded to $a\omega$, thus $E(X')=a\omega$ and $V(X')=0$. Figure 5b shows the

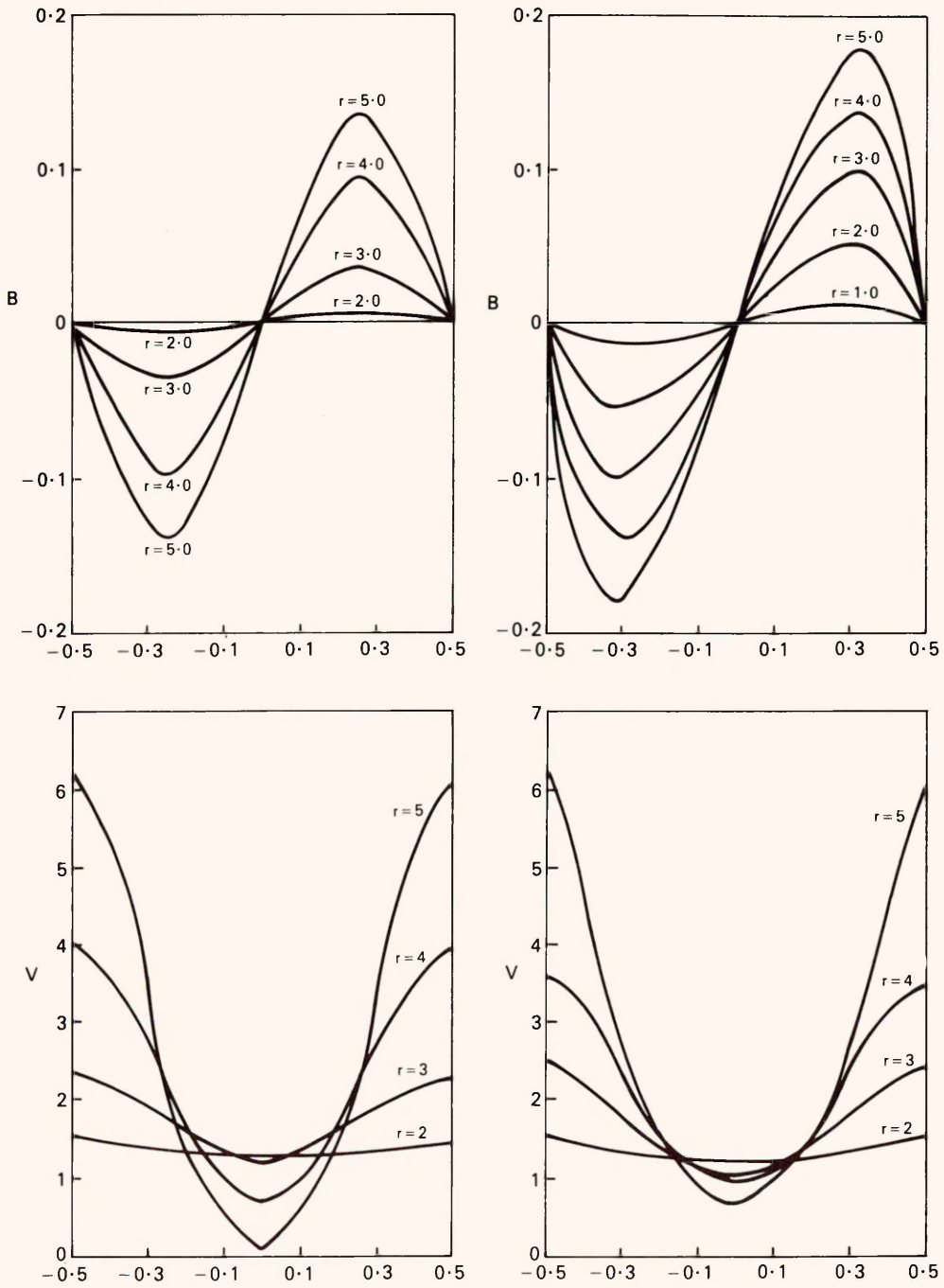


Fig. 1. *B* (equation (3.5)) for values of *a* between ±½ and *r* ranging up to 5 for normal distribution.
Fig. 2. *B* (equation (3.5)) for values of *a* between ±½ and *r* ranging up to 5 for Laplace distribution.
Fig. 3. *V* (equation (3.6)) for values of *a* between ±½ and *r* ranging up to 5 for normal distribution.
Fig. 4. *V* (equation (3.6)) for values of *a* between ±½ and *r* ranging up to 5 for Laplace distribution.

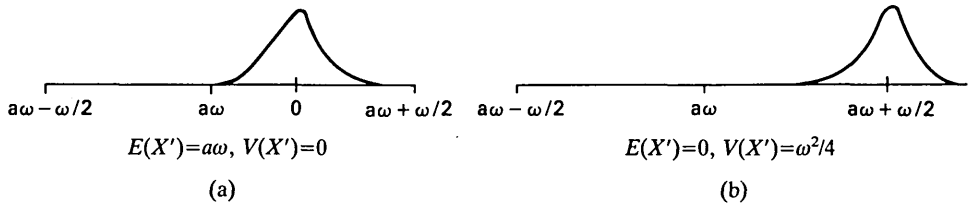


Fig. 5. Rounding of values x with normal and Laplace distribution when r is large. (a) When zero does not coincide with a cell boundary. (b) When zero does coincide with the cell boundary $a\omega+\omega/2$.

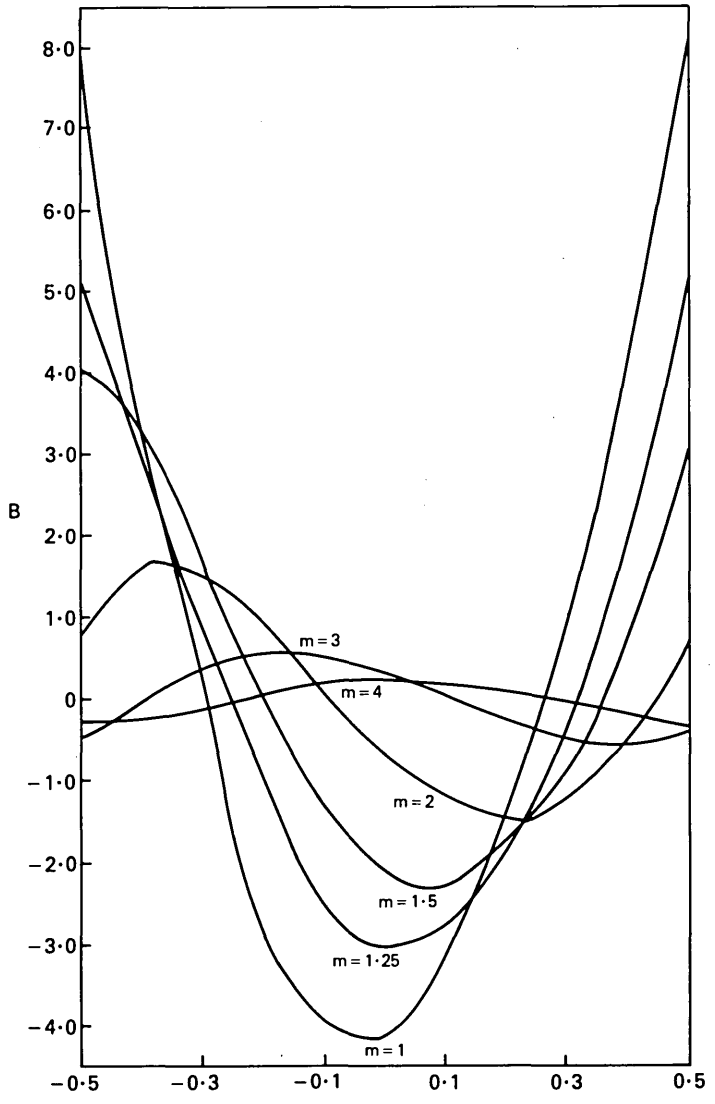


Fig. 6. B (equation (3.5)) for values of a between $\pm\frac{1}{2}$, $r=1$ and m up to 4 for gamma distribution.

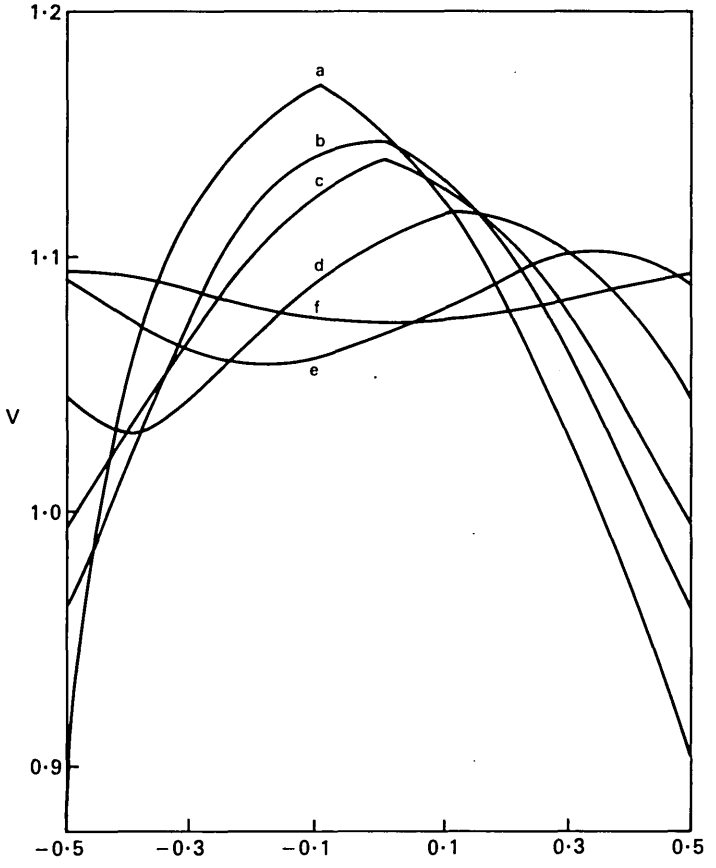


Fig. 7. V (equation (3.6)) for values of a between $\pm\frac{1}{2}$, $r=1$ and m up to 5 for gamma distribution.

Table 1. Maximum errors expected for mean caused by rounding (percentage of standard deviation)

m	r					
	0·25	*0·5*	*1·0*	*1·5*	*2·0*	*3·0*
1·0	0·5[a]	2·1	8·2	18·1	31·4	65·6
1·25	0·2	1·1	5·2	12·8	23·9	55·5
1·50	0·1	0·6	3·1	8·7	17·9	46·5
1·75	$0{\cdot}5\,(10)^{-1}$	0·4	2·3	6·8	14·0	38·7
2·0	$2{\cdot}5\,(10)^{-2}$	0·2	1·6	5·5	12·5	34·5
3·0	$2{\cdot}7\,(10)^{-3}$	$4{\cdot}1\,(10)^{-2}$	0·6	2·7	7·3	26·4
4·0	$3{\cdot}3\,(10)^{-4}$	$9{\cdot}6\,(10)^{-3}$	0·3	1·7	5·3	22·1
5·0	$1{\cdot}3\,(10)^{-4}$	$2{\cdot}6\,(10)^{-3}$	0·1	1·1	4·2	20·0

[a]Indicates maximum error in mean is 0·5 per cent of σ.

situation when the origin of the distribution coincides with the cell boundary $a\omega+\omega/2$. In this situation $a=-\frac{1}{2}$. Half the values will be rounded to $a\omega$ and half to $a\omega+\omega$. Thus $E(X')=a\omega+\omega/2=0$ and $V(X')=\omega^2/4$.

4 Non-symmetrical distributions – gamma

The gamma distribution which has the following probability density:

$$f(x)=\frac{\lambda}{\Gamma(m)}(\lambda x)^{m-1}\,e^{-\lambda x}\quad x\geq 0,\qquad \lambda>0,\qquad m\geq 1 \tag{4.1}$$

with mean m/λ and variance $\sigma^2=m/\lambda^2$.

Using the characteristic function of (4.1), we can obtain from (2.7) and (2.8) expressions for the first two moments in terms of r.

$$E(X')=E(X)+\sigma\left\{\frac{r}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}G^{-m/2}\sin(m\theta-2\pi ka)\right\} \tag{4.2}$$

$$E(X'^2)=E(X^2)+\sigma^2\left[\frac{r^2}{12}+\frac{r^2}{\pi}\sum_{k=1}^{\infty}(-1)^k\,G^{-m/2}\left\{\frac{1}{\pi k}\cos(m\theta-2\pi ka)+\frac{2}{r}\left(\frac{G}{m}\right)^{-1/2}\right\}\sin\{(m+1)\,\theta-2\pi ka\}\right] \tag{4.3}$$

where

$$\tan\theta=\frac{2\pi k}{r\sqrt{m}},\qquad G=\left\{1+\frac{(2\pi k)^2}{r^2 m}\right\}$$

For a given value of r, the effect of rounding on the moments is determined by the skewness of the distribution. For example, equation (4.2) shows that this is true for the mean. $E(X')$ is dominated by the factor $G^{-m/2}$, which for fixed r approaches zero as m increases. Thus the rounding process causes less bias in the mean as the gamma distribution becomes more symmetrical.

Figures 6 and 7 show curves for B and V for a ranging between $-\frac{1}{2}$ and $\frac{1}{2}$, $r=1$ and m up to 5. For various values of m and r the maximum errors expected for the mean and variance

Table 2. Maximum errors expected for variance caused by rounding (percentage of variance)

m	r					
	0·25	*0·5*	*1·0*	*1·5*	*2·0*	*3·0*
1·0	1·1[a]	4·1	17·5	40·0	72·0	171·0
1·25	0·8	3·5	15·1	35·9	65·2	160·8
1·50	0·7	3·1	11·4	33·6	64·0	160·1
1·75	6·1 $(10)^{-1}$	2·9	11·3	32·3	61·0	156·5
2·0	5·9 $(10)^{-1}$	2·6	11·2	30·5	60·3	153·8
3·0	5·3 $(10)^{-1}$	2·2	9·7	27·0	54·7	149·9
4·0	5·2 $(10)^{-1}$	2·1	9·4	23·6	52·0	146·6
5·0	5·2 $(10)^{-1}$	2·0	8·9	23·3	49·6	143·9

[a]Indicates maximum error in variance is 1·1 per cent of σ^2.

Table 3. Maximum expected for mean and variance caused by rounding for the gamma and normal distribution

m	*Maximum error in mean (percentage of standard deviation)*			*Maximum error in variance (percentage of standard deviation)*		
	r			r		
	1·1	2·1	3·0	1·0	2·0	3·0
10	$1{\cdot}1\ (10)^{-2}$	2·1	15·5	8·4	43·1	139·3
20	$6{\cdot}0\ (10)^{-4}$	1·2	12·6	8·3	39·6	132·7
30	$1{\cdot}0\ (10)^{-4}$	$8{\cdot}7\ (10)^{-1}$	12·3	8·3	38·5	132·5
40	$5{\cdot}2\ (10)^{-7}$	$7{\cdot}7\ (10)^{-1}$	11·9	8·3	38·1	132·0
Normal	$8{\cdot}1\ (10)^{-8}$	$4{\cdot}4\ (10)^{-1}$	10·1	8·3	36·7	130·0

are given in Tables 1 and 2 respectively. The influence of skewness can be seen in Figure 6. As m increases, the range in bias decline. For the exponential distribution ($m=1$) the bias in the first two moments is most severe. As m increases the bias quickly reduces. For $r=1$ the maximum error in the mean is $8{\cdot}2\sigma(10)^{-2}$ at $m=1$, while $m=2$ it is $1{\cdot}6\sigma(10)^{-2}$. As expected for increasing m, the errors in the moments approach those for the normal distribution (Table 3). Generally the $V(X')$ is greater than $V(X)$. In the outer ranges of a, the value of V is less than 1, indicating that $V(X')$ is less than $V(X)$. This has implications for estimation procedures, which have been discussed in Tricker (1984) for $m=1$.

5 Conclusions

By obtaining the characteristic function of the rounded random variable X, we have found a method of determining how the rounding process affects the moments. The differences between the moments of X and X' depend on three main factors: the skewness of X, the degree of rounding (r) and the position of the rounding lattice (a).

The degree of skewness of a distribution is of crucial importance in determining how the moments of a distribution can be distorted by rounding. For symmetrical distributions, such as the normal and Laplace, the maximum errors in the mean and variance are small even for $r=2$. However, when the distribution is highly skewed, the situation changes. For the exponential distribution the errors in the mean and variance can be considerable for $r>0.25$. Results from the gamma distribution illustrate that when the degree of skewness reduces so does the error in the moments.

Generally as r decreases, so does the effect of rounding. The value of r for which the errors in the mean and variance are small depends on the skewness of the distribution. For example the errors for the gamma may be considered negligible for $m=1$, $r<0{\cdot}25$, and for $m=2$, $r<0{\cdot}5$.

The influence of the position of the rounding lattice on the moments is less important than both the skewness of the distribution and the value of r.

When data are rounded, it is often assumed that no account need be taken of the resulting error. This is reasonable for the normal distribution but may not be sensible for skewed distributions. Previous work on the precision of data has concentrated on the effect of the degree of precision of the recorded data (r) on the distribution. Our analysis and examples suggest that this is not the only important factor. The position of the rounding lattice, and especially the skewness of a distribution must be taken into account.

Acknowledgements

The author would like to thank Dr G. K. Kanji, Dr D. A. Preece and Dr D. N. Shanbhag for their helpful suggestions in the preparation of the paper.

References

Eisenhart, C. (1947). Effects of rounding on grouped data. *Selected Techniques of Statistical Analysis* (ed. C. Eisenhart, M. Hastay and W. A. Wallis), Chapter 4. New York.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A,* **222,** 309–68.

Gjedderbaek, N. E. (1968). Grouped observations. *International Encyclopaedia of Social Sciences,* (ed. D. L. Sills) Vol. 15, pp. 193–6. Macmillan.

Kendall, M. G. and Stuart, A. (1968). *The Advanced Theory of Statistics,* Vol. 1. Griffin, London.

Lowell, M. S. (1980). On round-off error. *Analytical Chemistry,* **52,** 1141–7.

Tricker, A. R. (1984). Effects of rounding data sampled from the exponential distribution. *Journal of Applied Statistics,* **11** (1), 51–87.

Watts, D. G. (1961). A general theory of amplitude quantisation with applications to correlation determination. *Proceedings IEE, ptC,* **109,** 209–18.

Journal of Applied Statistics, Vol 11, No. 1, 1984

EFFECTS OF ROUNDING DATA SAMPLED FROM THE EXPONENTIAL DISTRIBUTION

Tony Tricker
Department of Mathematics, Statistics & Operational Research
Sheffield City Polytechnic, Sheffield, UK

Abstract

This paper looks at the effects of rounding data sampled from the exponential distribution. It examines the nature of the rounded distribution, together with the resulting error distribution. The influence of these distributions on estimates and tests of hypothesis is investigated. The results indicate that even a moderate degree of rounding can cause the bias in an estimator to increase, whereas in hypothesis tests level of significance is altered.

1. Introduction

When dealing with data, we are usually forced to round our values to a certain degree of precision. This introduces "rounding error", the size of which can have an important influence on the statistical influences that are to be made. Some of the consequences of rounding errror were discussed by Eisenhart (1947), Fisher (1922), Gjedderbeak (1968), Kulldorff (1961) and Lowell (1968). Most of these works are confined to the normal distribution and only Kulldorff (1961) dealt with the exponential distribution, when he showed how the method of maximum liklehood can be used for grouped data. Very little has been written on how rounding error can distort the information conveyed by data drawn from an exponential population. The effect of this distortion on the mean and variance is shown, with its possible consequences on estimation and hypothesis testing.

Eisenhart's (1947) recommendation that the width of the rounding interval should be either less than one third or one fourth of the standard deviation is also considered.

If values from a random variable X are rounded, the result is a new random variable X'. Let x and x' represent values from the random variable X and X' respectively. Then we may write:

$$x' = x + e$$

where e is the rounding error, itself a random variable which will be denoted by Z. If w is the width of the rounding interval, so that any value between $x-w/2$ and $x+w/2$ will be rounded to x', then Z will be distributed between $-w/2$ and $w/2$. Often it is assumed that:

(i) Z is uniformly distributed on (-w/2, w/2)

(ii) X and Z are independent.

The implications of these assumptions are:

(iii) $E(Z) = 0$, $V(Z) = w^2/12$, which may imply that

(iv) $E(X') = E(X)$; $V(X') = V(X) + w^2/12$.

In fact X and Z cannot be treated as independent and hence the validity of the above statements are in doubt. However the statements (i), (iii) and (iv) may be suitable in many practical situations. The validity of these three statements, together with the general effect of rounding on the expon ential distribution will be looked at in this paper.

2. Distribution of X' the rounded variable

If the random variable X follows an exponential distribution, then the probability density function is:

$$f(x) = \frac{1}{\theta}e^{-x/\theta} \qquad x \geqslant 0, \theta \geqslant 0$$

where $E(X) = \theta$ and $V(X) = \theta^2$.

If values of x are rounded into intervals of width w with midpoints x', and the centre of the interval containing zero is c, then x' has the following values:

$$c, c+w, c+2w, c+3w, \ldots \qquad (2.1)$$

(2.1) is known as the rounding lattice. The probability distribution of X' is:

$$P(X' = mw+c) = \begin{cases} \int_0^{w/2+c} f(x)\,dx & \text{for } m = 0 \\ \int_{mw+c-w/2}^{mw+c+w/2} f(x)\,dx & \text{for } m = 1,2,3.... \end{cases} \qquad (2.2)$$

X' is the rounded distribution and Figure 1 illustrates how its distribution is formed.

The probability distribution of X' from (2.1) is:

$$P(X' = mw+c) = \begin{cases} 1-\varepsilon^{1/\theta(c+w/2)} & \text{for } m = 0 \\ (e^{w/2\theta}-e^{-w/2\theta})e^{(-1/\theta)(mw+c)} & \text{for } m = 1,2.... \end{cases} \qquad (2.3)$$

Where c is the centre of the interval containing zero and its value determines the position of the rounding lattice on the underlying distribution. Often the rounding lattice is imposed at random on the underlying distribution. This means when values from an exponential distribution are rounded, zero shall not inevitably be the lower extremity of the lowest rounding interval - in fact c itself may be a negative number. Thus c may be located at random between $-w/2$ and $w/2$.

If we let $w = r\sigma$, then r measures the degree of rounding with respect to the standard deviation σ. As r indicates the severity of rounding it will be useful to express the probability distribution of X' given in (2.3) in form of r. Thus we have:

$$P(X' = (m+a)w) = \begin{cases} 1-e^{-r(a+\frac{1}{2})} & \text{for } m = 0 \\ (e^{r/2}-e^{-r/2})e^{-r(m+a)} & \text{for } m = 1,2,3.... \end{cases} \qquad (2.4)$$

where $r = w/\sigma = w/\theta$ and $a = c/w$. In fact a lies between $-\frac{1}{2}$ and $\frac{1}{2}$, and it determines the position of c. For example $a = \frac{1}{4}$ gives $c = w/4$, thus the centre of the interval containing zero is $w/4$. The cumulative distribution of X' for $r = 2, 1, \frac{1}{2}, \frac{1}{4}$ and $a = 0$ or $\frac{1}{2}$, are given in Figures 2, 3, 4, 5 and 6 respectively, where they are compared with the corresponding distribution of X, which is exponential.

The large number of steps in the X' distribution shown in these figures, illustrates its discontinuous nature. It is impossible to find smooth curves that will approximate these step functions closely at all points. The $P(X \leqslant w)$ will approximate closely the $P(X' \leqslant w)$ only for certain values of w, ie those near a point of intersection of the X curve with the horizontal position of a step in X'. This approximation improves considerably as the value of X' increases, being caused by the tail off effect of the exponential distribution.

3. Mean and Variance of X'

Using the probability distribution of X' given in (2.4) we have:

$$E[X'] = \theta\left(ar + \frac{re^{ar}}{e^{r/2}-e^{-r/2}}\right) \tag{3.1}$$

$$V(X') = \theta^2\left(r^2 e^{-r(3/2+a)}\,\frac{[1+e^{r}-e^{r(\frac{1}{2}-a)}]}{(1-e^{-r})^2}\right) \tag{3.2}$$

Of particular interest is the difference between E[X'] and E[X] and between V(X) and V(X'). Equations (3.1) and (3.2) can be rewritten as follows:

$$M = \frac{E[X'] - E[X]}{E[X]} \times 100 \tag{3.3}$$

$$V = \frac{V[X'] - C(X)}{V(X)} \times 100 \tag{3.4}$$

Expressions (3.3) and (3.4) now represent the percentage changes in E[X] and V(X) caused by rounding. Figures 7 and 8 show curves for M and V respectively for a ranging between $-\frac{1}{2}$ and $\frac{1}{2}$, and r up to 1. Figure 7 shows that the value of M is influenced considerably by the value of r. For a fixed a, as r increases the departure of M from zero becomes greater. On this scale for $r \leqslant \frac{1}{4}$ the curves are indistinguishable from a horizontal line passing through M = 0. For r = 1, M ranges between -4.05 to 8.20. As expected the range in M decreases as r decreases. Of interest is the region in Figure 7 where the curves intersect the line M = 0, indicating

$E[X] = E[X']$. These two regions are in around a = -0.3 and a = 2.5. This could be of relevence in the estimation of θ, dealt with in Section 5.

Figure 8 again shows the influence of r and a on V. For $r = \frac{1}{4}$, the range in V is only -0.52 to 1.06. Further calculations show that for $r \leqslant \frac{1}{4}$, the maximum values of M and V are 0.5 and 1.1 respectively. It is thus reasonable to assume that $E[X'] = E[X]$ and $V[X'] = V[X]$ for $r \leqslant \frac{1}{4}$.

Of interest is the limiting behaviour of the expectation and variances of X'. We first consider what happens when r approaches zero, ie when the rounding interval w is much smaller than θ. From equations (3.1) and (3.2) it can be shown that as r approaches zero, E(X') and V(X') tend to θ and θ^2 respectively. When r approaches ∞ then w is much larger thanθ. In this situation the position of c is very important. Figure 9a shows when the exponential distribution is between the cell boundaries and zero does not occur on a cell boundary. All values of x will be rounded to c, thus $E[X'] = c$. As all the values of X' are the same then $V(X') = 0$. Figure 9b shows the situation when the origin of the exponential distribution coincides with the cell boundary $c + w/2$. In this situation the value of c is $-w/2$. All the x values will be rounded to $c + w$ which equals $w/2$; giving $E[X'] = w/2$. Again as all the values of X' are the same the $V(X') = 0$. When zero coincides withthe other cell boundary $c-w/2$, then $E[X'] = c = w/2$.

4. The Error Distribution

As the distribution of errors is often assumed to be uniform on $[-w/2, w/2]$, it will be of interest to investigate whether this assumption can reasonably be made for the **exponential** distribution.

If the probability distribution of the error distribution is denoted by g(z) then:

$$g(z) = \begin{cases} \sum_{k=0}^{\infty} f(c + kw - z) & -w/2 \leq z \leq c \\ \sum_{k=1}^{\infty} f(c + kw - z) & c \leq z \leq w/2 \end{cases} \qquad (4.1)$$

where f(.) is the exponential probability density function. We may express g(z) as:

$$g(z) = \begin{cases} \left[\frac{e^{w/\theta}}{e^{w/\theta}-1}\right] \frac{1}{\theta} e^{-1/\theta(c-z)} & -w/z \leq z \leq c \\ \left[\frac{1}{e^{w/\theta}-1}\right] \frac{1}{\theta} e^{-1/\theta(c-z)} & c \leq z \leq w/z \end{cases} \qquad (4.2)$$

The distribution g(z) is shown in Figure 10. The value of c is important in determining this distribution, as can be seen from equation (4.2). Often, the rounding error of rounded data is assumed to follow a uniform distribution. Quite obviously g(z) is far from being a flat topped distribution. To investigate the departure of g(z) from uniformity we can consider the following function:

$$h(z) = \frac{g(z) - 1/w}{1/w} \times 100 \qquad (4.3)$$

We shall consider the value of this function for values of $z = (-0.5 + 0.05k)w$ for $k = 1, 2, \ldots, 21$. It can be shown that the value of (4.3) is determined by r and c. As r is the main influence on the value of h(z) we shall only consider its value for $c = 0$. Table 1 contains the values of h(z) for various r values. As expected for $r \geq 1$, the departure from uniformity is considerable. As r decreases the g(z) distribution tends to uniformity.

From (4.2) we have:

$$E[Z] = -\theta + c + \frac{we^{-w/\theta}}{e^{w/2\theta} - e^{-w/2\theta}} \qquad (4.4)$$

Rewriting (4.4) in terms of r and a we have

$$E[Z] = \theta\left[ar + \frac{re^{-ar}}{e^{r/2} - e^{-r/2}} - 1\right] \qquad (4.5)$$

As expected E[Z] approaches zero as r tends to zero. However the value of (4.5) for specific r's and a's is important as it is often assumed that E[Z] = 0. Table 2 gives the end points of the possible range in values of E[Z] for given values of Z, the end points being expressed as percentages of the parameter θ. The wide range in values is caused by the lattice effect.

5. Estimation of

Given a random sample of size n drawn from the exponential distribution, the unbiased estimate of θ is $\bar{X}$. However, if the data has been rounded what influence will this have on the estimator? Using (3.1) we have

$$E[X'] = \theta\{ar + \frac{re^{-ar}}{e^{r/2}-e^{-r/2}}\} \tag{5.1}$$

$$V(X') = \theta^2\{r^2 e^{-r(3/2+a)}[\frac{1+e^{r}-e^{r(\frac{1}{2}-a)}}{(1-e^{-r})^2}]\} \tag{5.2}$$

As $E[\bar{X}']$ is independent of n, the bias caused by rounding namely $E[\bar{X}']-\theta$ does not decrease to zero even if n becomes large. On the other hand, the bias in $V(\bar{X}')$, namely $V(\bar{X}') - \frac{\theta^2}{n}$ does contain n, and so vanishes for large n. If the data are rounded $\bar{X}'$ is no longer on unbiased estimates of θ. This bias in $\bar{X}'$ depends on r and on the position of the rounding lattice. We can use the following expression to express this bias relative to the rounding interval w.

$$B = \frac{E[X'] - \theta}{W} \tag{5.3}$$

Figure 11 shows the curves of B for various values of r ranging up to 2. For r less than 0.1 the curves are almost horizontal passing through zero. Rounding error has caused $\bar{X}'$ to become a biased estimator of θ. As expected the biasness increases as rounding becomes more severe. The position of the rounding lattice has considerable influence on how biased an estimator $\bar{X}'$ can become. For $r < 2$ the bias is minimum

in the region of a = -0.3 and a = 0.25. This implies that to reduce the bias in $\bar{X}'$ as an estimator of θ, we should choose the centre of the interval containing zero in the region of -0.3w and 0.25w.

We can show the effect of rounding on the variance by the following expression:

$$R = \frac{V(\bar{X}')}{V(\bar{X})} \qquad (5.4)$$

Figure 12 shows the curves for R, for r ranging up to 2. For r less than 0.25 the curves are almost horizontal passing through 1. Generally as r increases there is a corresponding increase in R, indicating that the $V(\bar{X}')$ is greater than $V(\bar{X})$. However in the outer ranges of a, the value of R is less than 1, indicating that the $V(\bar{X}')$ is less than $V(\bar{X})$. This implies that rounding may sometimes cause $\bar{X}'$ to be more precise than $\bar{X}$ as an estimator of θ, although, rounding has resulted in a reduced amount of information from the data.

However, we have only considered the $E[\bar{X}']$ and $V[\bar{X}']$ for $r \leqslant 2$. Using the results from Section 3, it is easy to see that the $E[\bar{X}']$ will tend to a limit and $V(\bar{X}')$ will tend to zero for large r.

Whether an estimate is unbiased is not the only criteria, the size of the sampling variance is also important. Let us, therefore, consider the mean-square-error (M.S.E.) of an estimator.

$$\text{M.S.E.} = E[\hat{\theta} - \hat{\theta}]^2 = V(\hat{\theta}) + [E(\hat{\theta} - \hat{\theta})]^2$$

where $\hat{\theta}$ is the estimator of $\hat{\theta}$.

Obviously if $\hat{\theta} = \bar{X}$ the M.S.E. is $V(\bar{X})$, which is $\frac{\theta^2}{n}$. If we round the data and use $\hat{\theta} = \bar{X}'$ then

$$\text{M.S.E.} = V(\bar{X}') + [E(\bar{X}') - \theta]^2 = \theta^2 b \qquad (5.5)$$

where b can be found from (5.1) and (5.2) for a given a and r.

Most practical situations will involve $r \lesssim 1$. Figure 13a and 13b show curves of M.S.E. for r ranging upwards to 1, where the sample size is 5 and 20. As expected the value of the M.S.E. is influenced by the size of r and a. The range in the value of the M.S.E. increases considerably as r increases. When $r \lesssim 0.1$ the M.S.E. of $\overline{X}$ and $\overline{X}'$ are almost equal. Of particular interest is where the minimum value of the M.S.E. occurs. It can be shown that when $n \lesssim 15$ the minimum value is situated at $a = \frac{1}{2}$ and $a = -\frac{1}{2}$ on the rounding lattice. As n increases beyond 15 the dominating influence on the M.S.E. value is the amount of bias in $\overline{X}'$. This causes the minimum value to move towards the position on the lattice where the bias in $\overline{X}'$ is least. This is illustrated in Figure 13b,where an increase in the sample size to 20 has caused the minimum value for r = 1 to move to a = 0.41. Examination of the M.S.E. for large values of r is really of no practical importance. However, it is easy to see that as r becomes large, the range in the value of M.S.E. increases and its minimum value occurs on the lattice where the bias in $\overline{X}'$ is least.

Of crucial importance is the effect of rounding on the mean and variance of the distribution. Table 3 gives the maximum errors expected for the mean and variance caused by rounding. When r = 1, the maximum error in the mean to 8.2 per cent of θ (8.2 per cent of the standard deviation) and for the variance is 17.5 per cent of θ^2 (17.5 per cent of the variance). For the normal distribution, Widrow (1961) shows that the maximum errors in the mean and variance are $8.3(10)^{-8}$ per cent of the standard deviation and $1.1(10)^{-6}$ per cent of the variance respectively (for r = 1). He showed that even for r = 2 the errors in the mean and variance are small. However, for the exponential distribution these errors can be considerable for small r, as Table 3 shows.

The errors in the mean are the amount of bias in $\overline{X}'$ as an estimator of θ. For a given value of r the amount of bias will depend on the value of a; the position on the rounding lattice. As mentioned in the section concerning the mean and variance of

$\bar{X}'$, a value of a can be chosen such that $\bar{X}'$ is an unbiased estimator of θ. However, this causes an increase in the variance of $\bar{X}'$, especially for small values of n. For this kind of situation, the minimum M.S.E. estimator for θ may be better.

6. Compensation for Rounding Error

For data rounded to a given precision, a value of a, can be chosen such that the sample mean $\bar{X}'$ is an unbiased estimate of θ. From equation (5.1) we have:

$$E[\bar{X}'] = \theta\{ar + \frac{re^{-ar}}{e^{r/2}-e^{-r/2}}\}$$

For a given r, $\bar{X}'$ will be an unbiased estimate of θ if a is chosen such that

$$ar + \frac{re^{-ar}}{e^{r/2}-e^{-r/2}} = 1 \qquad (6.1)$$

Table 4 gives the value of a which make $\bar{X}'$ an unbiased estimate of .

The above illustrates the importance of choosing a suitable value of a to reduce the bias in $\bar{X}'$ as an estimator of θ. For example if an experimenter has decided to round the data where $r \leqslant 2$, then choosing a in the region 0.23 to 0.29 will considerably reduce the bias in $\bar{X}'$.

Now consider the problem where we have obtained $\bar{X}'$ and we would like to compensate for the rounding error to make the sample mean a better estimator of θ. We use an approach similar to that of Lovell (1980), who compensated for rounding error, when estimating parameters in the normal distribution. Rounding error can be compensated for by using the following equation:

$$E[X'] = aw + \frac{we^{-aw/\theta}}{e^{w/2\theta}-e^{-w/2\theta}} \qquad (6.2)$$

If the sample size is not too small, $\bar{X}'$ will be a precise estimate of $E[\bar{X}']$ and we can obtain an estimate of θ, namely $\hat{\theta}$ from

$$\bar{X}' = aw + \frac{we^{-aw/\theta}}{e^{w/2\hat{\theta}} - e^{-w/2\hat{\theta}}} \tag{6.3}$$

Rewriting equation (6.3) we have

$$2(\bar{X}' - aw)\text{Sinh}[w/2] - we^{-aw/\hat{\theta}} = 0 \tag{6.4}$$

Obtaining $\hat{\theta}$ from equation (6.4) gives us an improved estimate of $\hat{\theta}$, in that the rounding error has been compensated. A solution to equation (6.4) can only be obtained only if w/θ is not too large. This restriction usually causes no problem. If the sample size is small then $\bar{X}'$ is a poor estimate of $E[\bar{X}']$, thus the rounding corrections themselves are random variables and subject to sampling error. This can cause $\hat{\theta}$ to be ineffective in compensating for the rounding error. For a = 0 equation (6.4) simplifies to:

$$\hat{\theta} = \frac{w}{2\text{Sinh}^{-1}(w/2\bar{X}')} \tag{6.5}$$

and for $a = \frac{1}{2}$

$$\hat{\theta} = \frac{w}{2\tanh^{-1}(w/2\bar{X}')} \tag{6.6}$$

For $a = \frac{1}{2}$, $\hat{\theta}$ is the maximum likelihood estimator of θ. Using the large sample properties of the maximum likelihood estimator we have

$$\hat{\theta} \sim N[\theta, \frac{4\theta^4\text{Sinh}^2(w/2\theta)}{nw^2}]$$

To illustrate how θ may compensate for rounding error, a set of data was simulated in the following way. A process was set up on the computer to generate a random sample of size n, of rounded data from the exponential distribution, with $\theta = 1$. The width of the rounding interval is w and a is set to zero. From this process the mean $\bar{X}'$ is calculated. Using equation (6.5) $\hat{\theta}$ was then calculated. Table 5 shows the results for n ranging between 50 and 1000 and r between 0.5 and 7. In Table 5,

column 3 shows the bias in $\overline{X}'$ and column 4 shows the bias in $\hat{\theta}$. Inspection of the table indicates that $\hat{\theta}$ is effective in giving an improved estimate for θ; for values of r between 1 and 3. For values of r less than 1, the sampling errors are large as compared with the bias in $\overline{X}'$; thus $\hat{\theta}$ becomes unreliable as an estimate of θ. For r greater than 3 the bias in $\overline{X}'$ increases considerably and as a result $\hat{\theta}$ tends to be very effective as an improved estimate of θ. Values for r = 7 in table 5 illustrate this point.

Although the bias in $\hat{\theta}$ may be less than $\overline{X}'$, what about the standard error (S.E.) of $\hat{\theta}$? Further simulation has shown that the S.E. of $\hat{\theta}$ is slightly smaller than that of $\overline{X}'$. When r = 1, the S.E. of is 0.96 per cent of the S.E. of $\overline{X}'$.

The above example shows that $\hat{\theta}$ is effective in compensating for rounding error for $r \geq 1$, where the sample size is 50 or more. As r increases in value, then $\hat{\theta}$ will be effective for smaller values of n.

7. Sampling Distribution of $S'n = \Sigma X'i$

When dealing with the exponential distribution we often require the sampling distribution of $Sn = \sum_{i=1}^{n} x_i$, when making inferential statements about the parameter θ. In this section we find the distribution of $S'n = \sum_{i=1}^{n} x'_i$ the sum of n rounded values from the exponential distribution.

A random sample $x_1, \ldots, x_n$ of size n is drawn from the exponential distribution. Each x_i is rounded into an interval of width w with midpoint x'_i, where the centre of the interval containing zero is aw. Then $S'n = \sum_{i=1}^{n} x'_i$ has the following probability distribution.

$$P[S'n = (m+an)w] = \begin{cases} (k')^n & \text{for } m = 0 \\ [A_{1n}k'^{n-1}ke^{-ra} + A_{2n}k(^{n-2}k^2e^{-2ra}(m-1)\ldots A_{nn}k^ne^{-nra}(m-1) \\ \quad \ldots.(m-n+1)]e^{-rm} & \text{for } m = 1,2,\ldots \end{cases}$$

$$P[S'n = (m+an)w] = \begin{cases} (k')^n & \text{for } m = 0 \\ e^{-rm}\sum_{j=1}^{n} A_{jn}k'^{n-j}k^je^{-jra}(m-1)(m-2)\ldots(m-j+1) & \text{for } m = 1,2,\ldots \end{cases}$$

(7.1)

where $k' = 1-e^{-r(\frac{1}{2}+a)}$, $k = e^{r/2}-e^{-r/2}$, $Ajn = \sum_{j}^{n} c/(j-1)!$, $r = w/\theta$ and $-0.5<a\leq 0.5$.

For a = -0.5 the probability distribution of S'n is the same as for a = 0.5.

Of interest is how well S'n compares with the continuous distribution Sn. The comparison may be made by investigating their cumulative distribution. The cumulative distribution for S'n can be obtained from equation (7.1). For Sn we have the relationship

$$2Sn/\theta \sim \chi^2_{2n} \tag{7.2}$$

or rewriting in terms of r and w

$$Sn \sim \frac{\chi^2_{2n} w'}{2r} \tag{7.3}$$

Thus the cumulative distribution of Sn can be expressed in terms of w for a given r and n.

Of interest is how well S'n compares with the continuous distribution Sn. In an earlier section we looked at the distribution of X', which is a special case of the S'n distribution where n = 1. Many of the observations we made concerning X' apply to S'n. Figures 14a to 14g show S'n for certain values of r, a and n, where it is compared with the appropriate distribution of Sn. As expected as r decreases in value the fit between S'n and Sn improves as r decreases. This is shown in figures 14a to 14d. Where the degree of rounding is severe, the disparity between the Sn and S'n distributions can be considerable, as shown in Figure 14a. As the size of the sample increases the magnitude of the steps in S'n decrease, in so doing improving the fit between S'n and Sn. This is illustrated in Figures 14b, 14e and 14f, where n is 5, 10 and 20 respectively. Although it is r not n which is the dominating factor in determining how close the fit is between S'n and Sn. Varying a, has the effect of shifting S'n. Figure 14g illustrates how changing a from 0 to 0.5 causes a shift to the right in the S'n distribution.

The distribution of Sn is important in determining confidence limits and carrying out testing hypothesis for θ. As a result the effect of the degree of rounding on the percentage points of Sn is crucial.

8. Effect of Rounding on the Percentage Points of Sn

Let S_1 represent that value of the Sn distribution with probability of obtaining values of Sn less than S_1 is α_1, or $P(Sn \leqslant S_1) = \alpha_1$. Similarly, let S_2 equal the point at which $P(Sn \geqslant S_2) = \alpha_2$. The values of S_1 and S_2 can easily be obtained by using equation (7.3). When the data is rounded we are dealing with S'n instead of Sn. Of interest is whether the value of α_1 and α_2 change as a result of the S'n distribution.

For a given α_1 and α_2 the points S_1 and S_2 on the Sn distribution are such that

$$P(Sn \leqslant \alpha_1) = \alpha_1 \qquad P(Sn \geqslant S_2) = \alpha_2$$

However the actual distribution is S'n, thus

$$P(S'n \leqslant S_1) = \alpha'_1 \qquad P(S'n \geqslant S_2) = \alpha'_2$$

Although we have chosen S_1 and S_2 according to the probabilities α_1 and α_2 on the Sn distribution, the actual probabilities will be α_1' and α_2' for rounded data. These may be obtained from the S'n distribution. To illustrate how rounding of the data affects the percentage points we shall consider the left hand tail of the Sn distribution, where α_1 has the values 0.05, 0.01 and 0.001. For these α values, α_1' was obtained for a fixed r and n, and a = -0.5, -0.4, ..., 0.5. Table 6 contains the range in values of α_1' for values of r and n. The range in the values has been caused by the position of the rounding lattice. We have only obtained the values of α_1' for only eleven positions on the rounding lattice. Thus the range given in the tables is an indication of the maximum possible range. To obtain the maximum range would have required considerable amount of computing, and this was not considered worthwhile. We have only considered samples up to size 25 only.

Table 6 indicates how rounding the data can cause the probability α_1 to change considerably. For example, S_1 is chosen for a sample of size 5 such that $\alpha_1 = 0.05$. Rounding the data will cause α_1 to change, as we are now dealing with the S'n distribution. For r = 1, α_1' can be between 0 to 0.102. The following example, illustrates how important the change in α_1 can be.

A random sample of size 5 is drawn from an exponential distribution with parameter . We wish to test the following hypothesis.

H_0: $\theta = 0.5$

H_1: $\theta < 0.5$

For $\alpha_1 = 0.05$ we will reject H_0 if $Sn \leq S_1$. Thus the probability of rejecting H_0 when in fact it is true is 0.05 (type 1 error). However, if the data is rounded, then α_1 is no longer 0.05, but may vary considerably. Suppose the data has been rounded, where the rounding interval w = 0.5, then r = 1 under H_0. In this situation α_1' can lie between 0 and 0.102. This means that the probability of a type 1 error can be anything between 0 to 0.102.

What determines the range of α_1'. The most important factor is the size of r. As expected, an increase in r will cause a corresponding increase in the range of α_1'. Even for r = 0.25, the difference between α_1 and α_1' can be considerable. For n = 5 and $\alpha_1 = 0.001$, α_1' can lie between 0 and 0.002. The influence of an increase in n on decreasing the range of α_1' can be seen from Table 6.

The problem is what combination of r and n is suitable such that the range in α_1' is tolerable. It appears from Table 6 that a satisfactory combination is:

$$\left.\begin{array}{ll} r \leq 0.1 & n \leq 15 \\ r \leq 0.25 & 15 < n \leq 25 \end{array}\right\} \qquad (8.1)$$

(8.1) is only an indication of the circumstances under which α_1 is not too greatly affected by rounding. More analysis is needed before a more detailed recommendation can be given.

Similarly we could consider the right hand tail of the Sn distribution. From the figures showing the cumulative distribution of Sn and S'n it can be easily seen that the fit between Sn and S'n is considerably improved in the right hand tail as compared with the left hand tail. Thus the percentage points in the right hand tail will be less influenced by rounding.

In this section we have only investigated the left hand tail of the Sn distribution. A more detailed analysis is required to formulate actual recommendations on the degree of rounding and its influence on the percentage points of Sn.

9. Conclusions

Rounding values of X always results in a discrete disribution X'. How well X' approximates to X depends on the degree of rounding (r) and the position of the rounding lattice (a). Reducing r from 1 to 0.25 (Figures 2, 3, 4 and 5) improves the fit between X and X' considerably. The influence of the value of a on this fit is less important than the value of r. Changing the value of a causes a shift in the distribution of X' (Figure 6).

The results in Table 1 indicate that the error distribution is definnitely not flat topped. For r = $\frac{1}{4}$ the departure from uniformity is still present. For any practical value of r it is unreasonable to assume that the error distribution is uniform.

The errors in the mean and variance of the exponential distribution may be considered negligible for $r \leq 0.25$. However for larger values of r these errors can cause the sample mean ($\bar{X}'$) to have a significant amount of bias, as an estimator of θ. Where the sample size is large enough it is possible to compensate for this bias.

Our analysis and examples suggest that Eisenhart (1947) is justified in recommending that the width of the rounding interval should be less than one fourth of the standard

deviation. However this degree of rounding is not precise enough for hypothesis testing for small sample sizes; the width of the rounding interval should then be less than one tenth of the standard deviation.

References

EISENHART C (1947). Effects of rounding on grouping data. Chapter 4 in Selected Techniques of Statistical Analysis (C. Eisenhart, M Hastay and W A Wallis eds) New York, McGraw Hill.

FISHER F A (1922). On the mathematical foundations of theoretical statistics. Phil Trans Roy Soc A, 222 pp 309-368.

GJEDDERBAEK N F (1968). Grouped observations pp 193-196 Vol 15 of D.L. Sills, ed, International Encyclopaedia of Social Sciences, Macmillan.

KULLODORFF GUNNAR (1961). Estimation from Grouped and Partially Grouped Samples, Almqvist and Wiksell, Stockholm.

LOWELL M S (1980). On Round-Off Error. Anal Chem 52, pp 1141-1147.

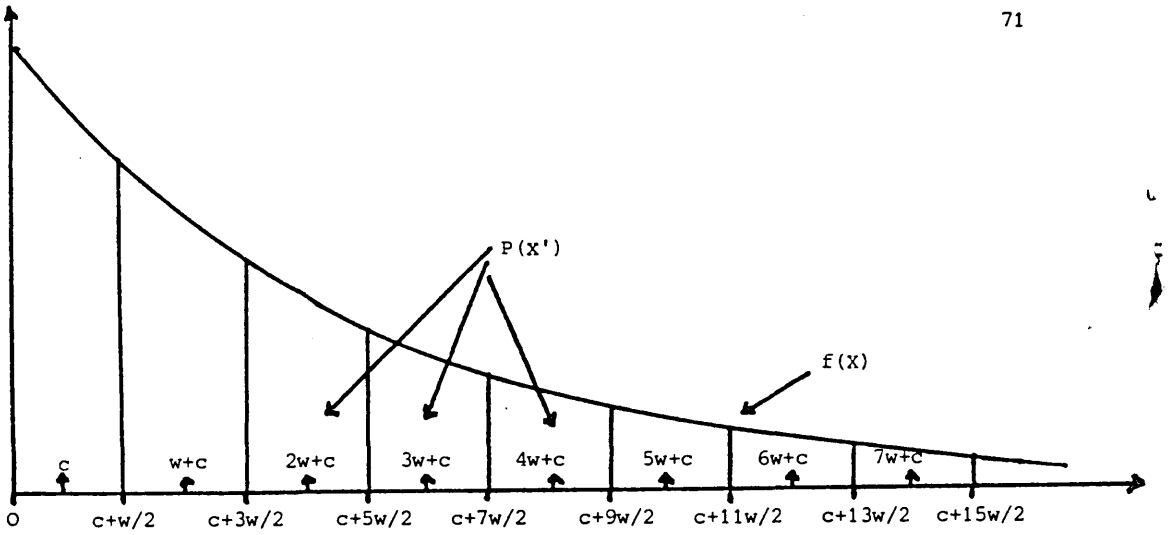


Figure 1. Derivation of x' distribution

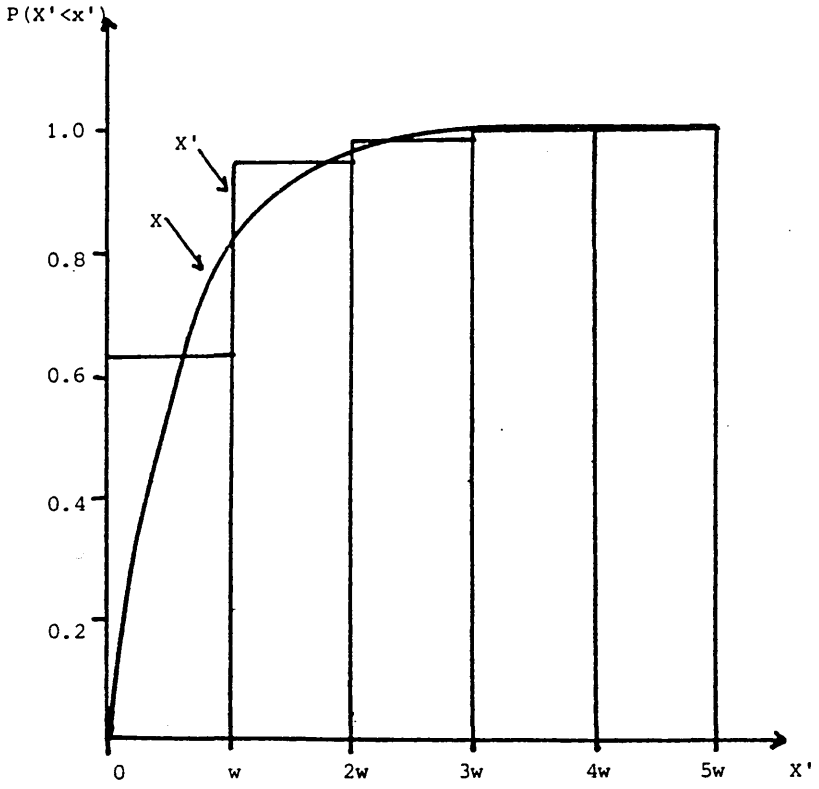


Figure 2. Cumulative distribution of X and X' where r = 2 and a = 0

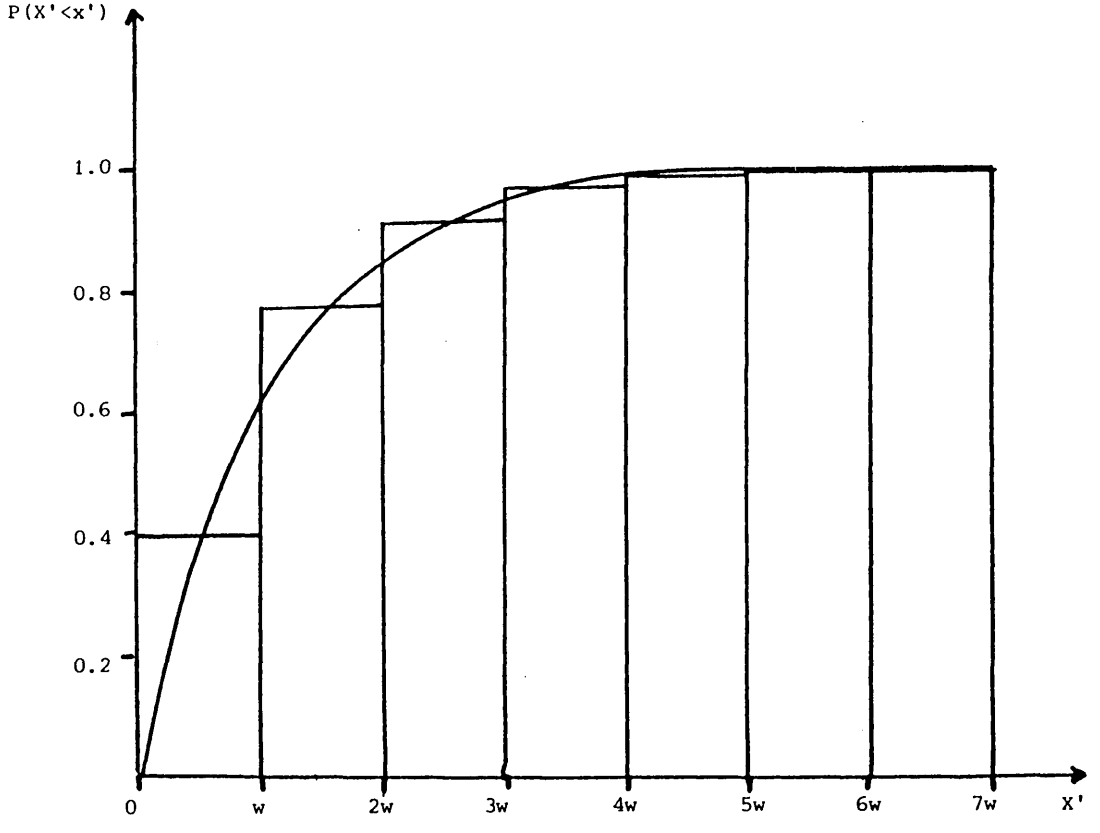


Figure 3. Cumulative distribution of X and X' where r = 1 and a = 0

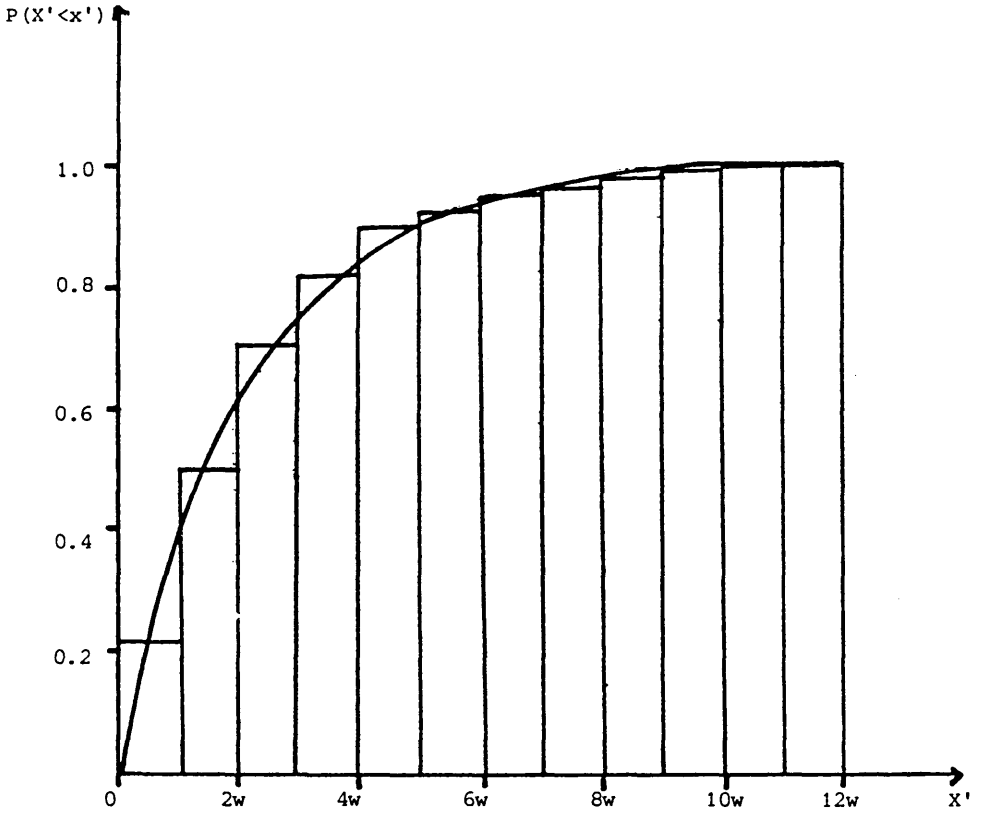


Figure 4. Cumulative distribution of X and X' where r = $\frac{1}{2}$ and a = 0

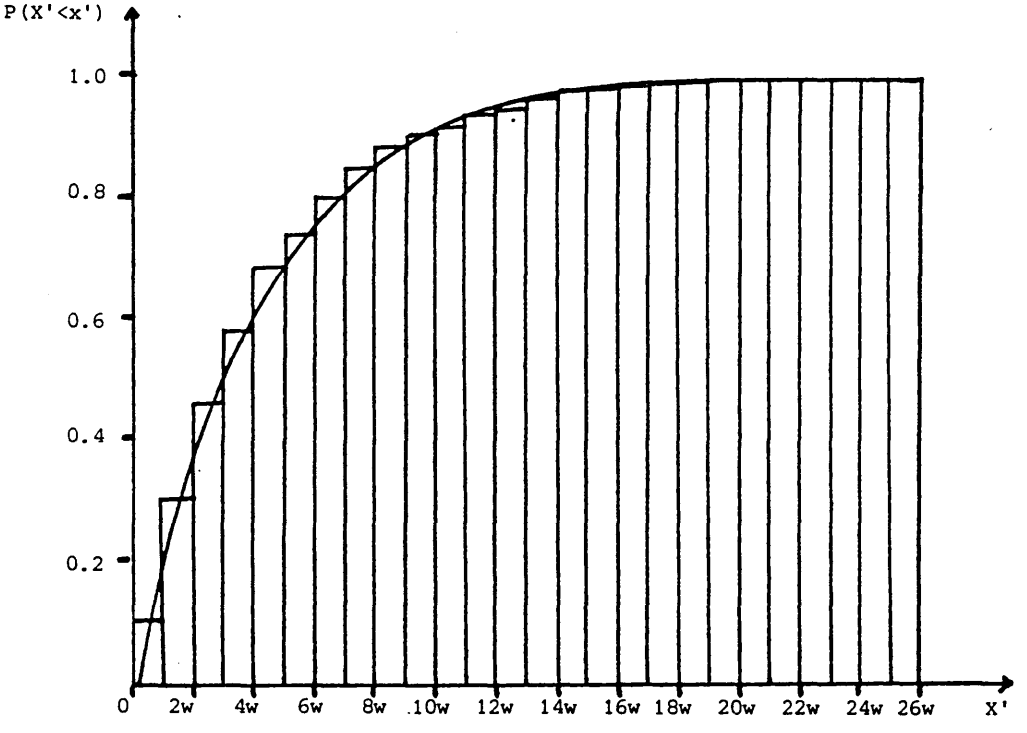


Figure 5. Cumulative distribution of X and X' where r = $\frac{1}{4}$ and a = 0

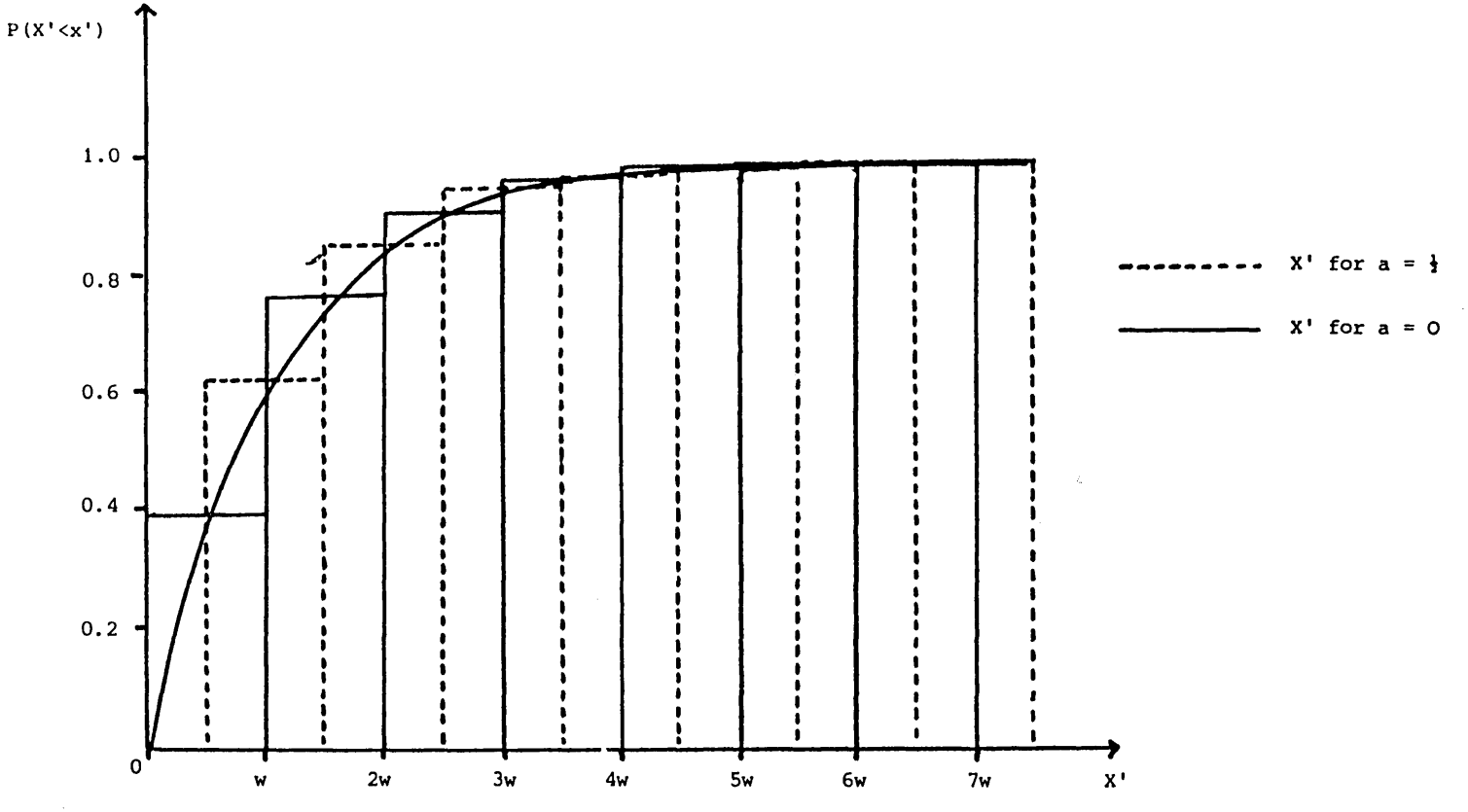


Figure 6. Cumulative distribution of X and X' where r = 1, a = 0 and a = ½

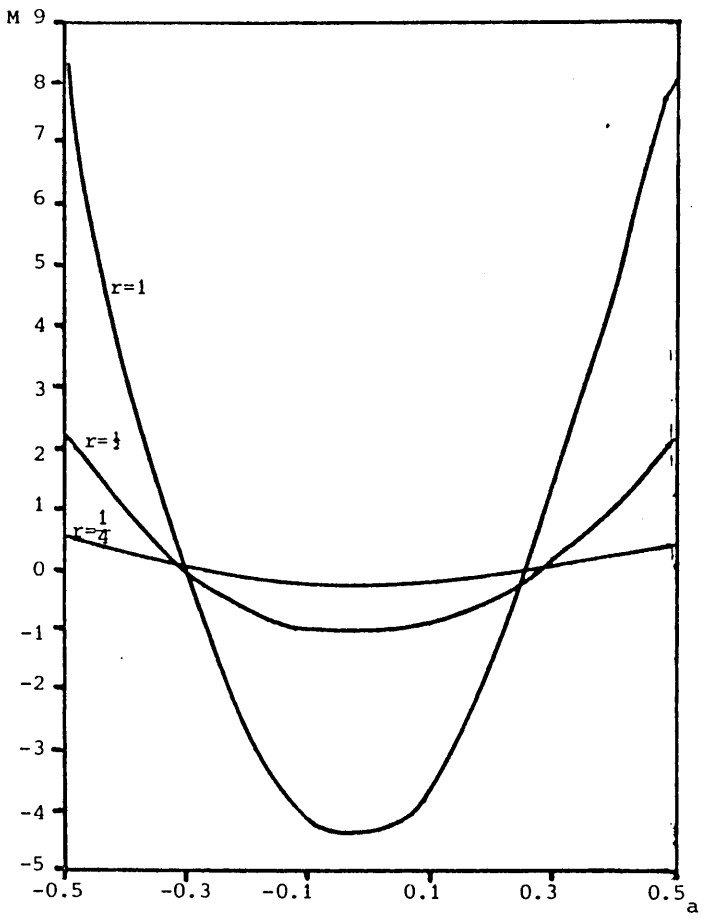


Figure 7. M(equation 3.3)for values of a between ±½ and r ranging up to 1.

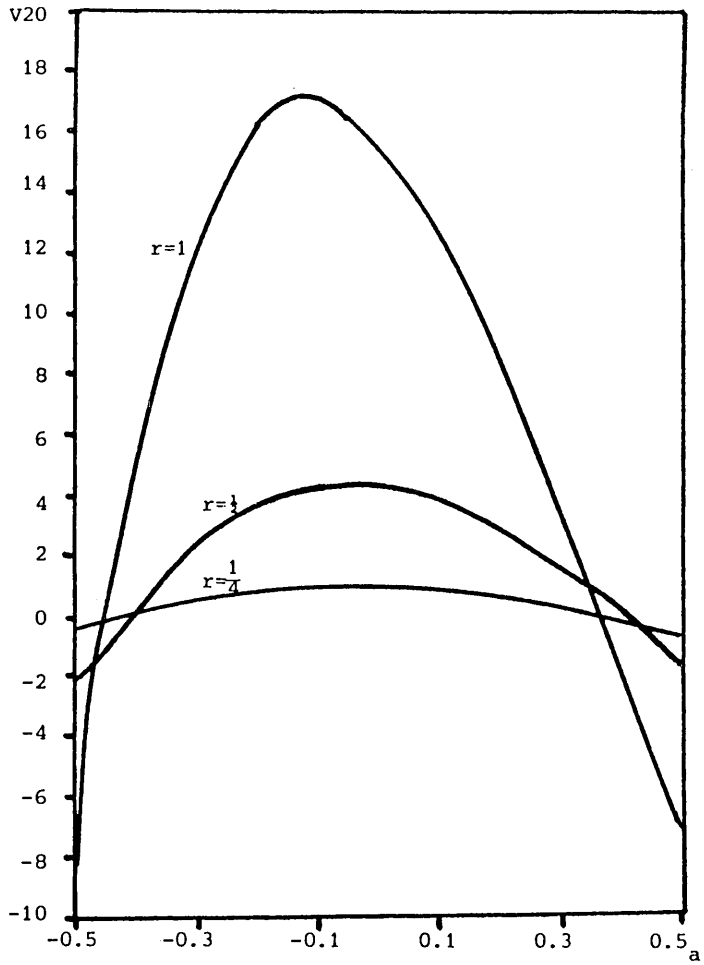
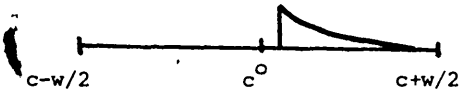
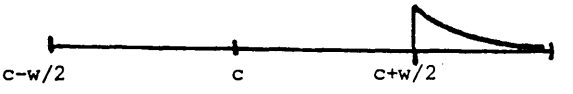


Figure 8. (equation 3.4) for values of a between ±½ and r ranging up to 1.



E(X') = c, V(X') = 0

Figure 9a.



E(X') = w/2, V(X') = 0

Figure 9b.

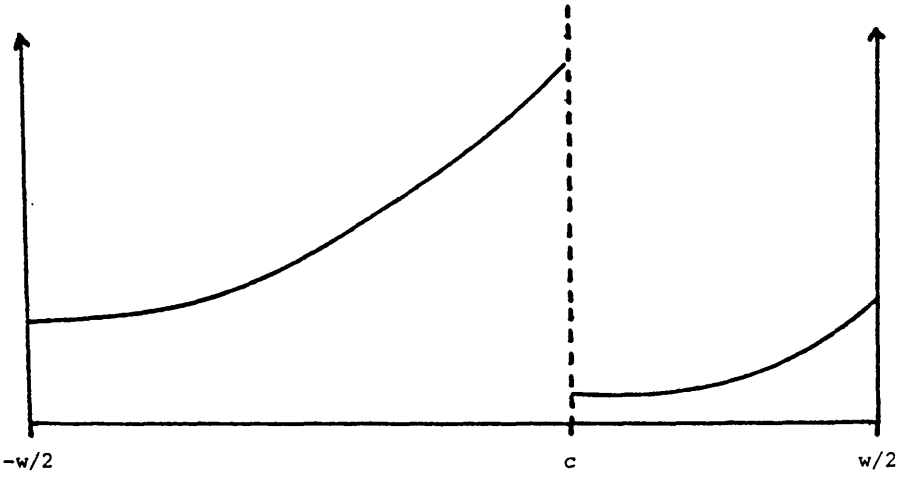


Figure 10. Error Distribution g(z)

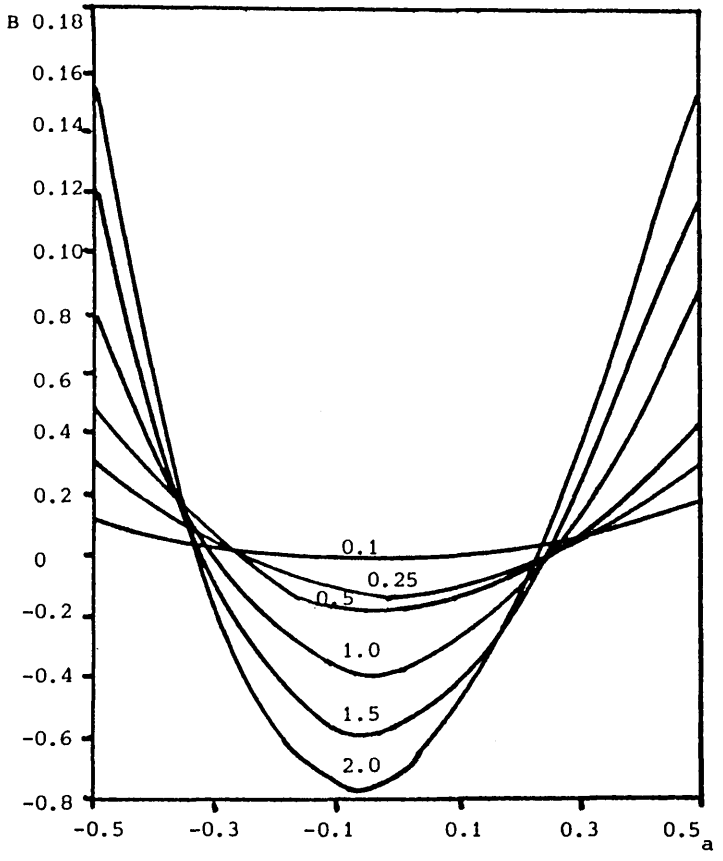


Figure 11. B(equation 5.3) for values of a between ±½ and r ranging up to 2

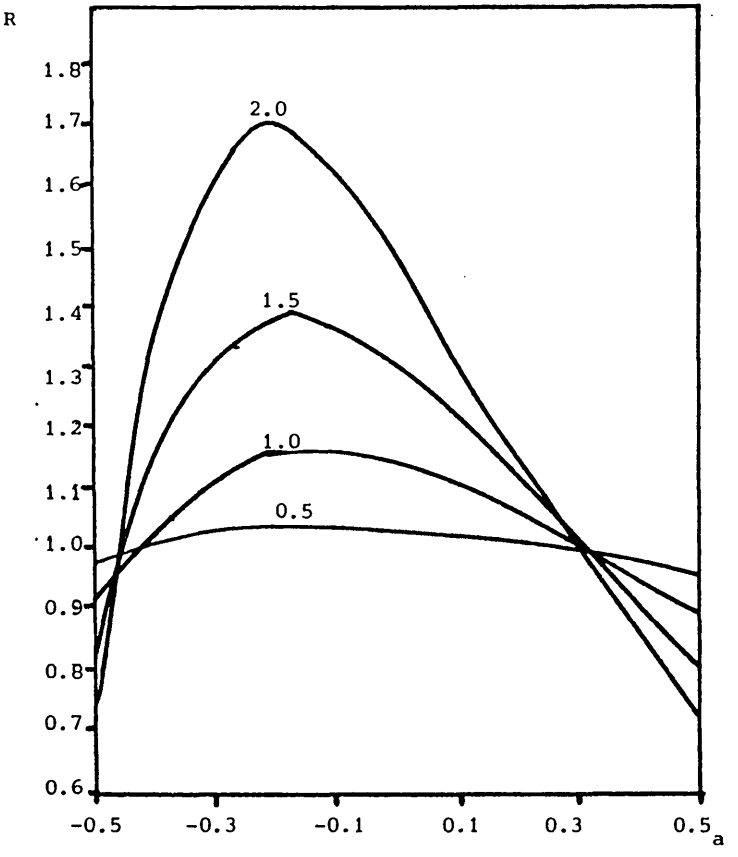
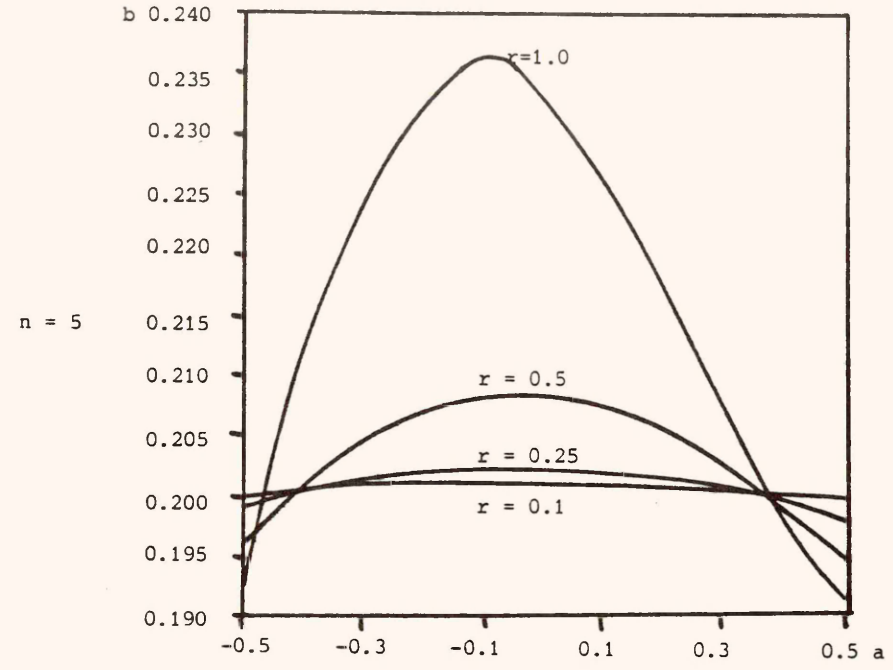
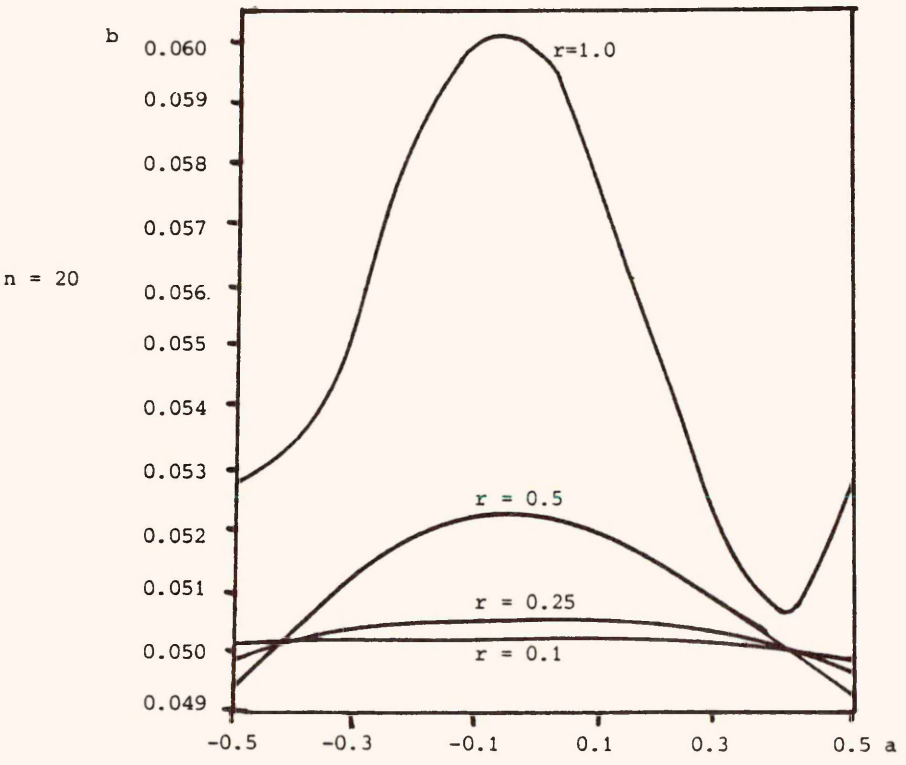


Figure 12. R(equation 5.4) for values of a between ±½ and r ranging up to 2





Figures 13a and 13b. MSE curves for n=5 and 20. Plotted for values of a between $\pm\frac{1}{2}$ and r ranging up to 1.

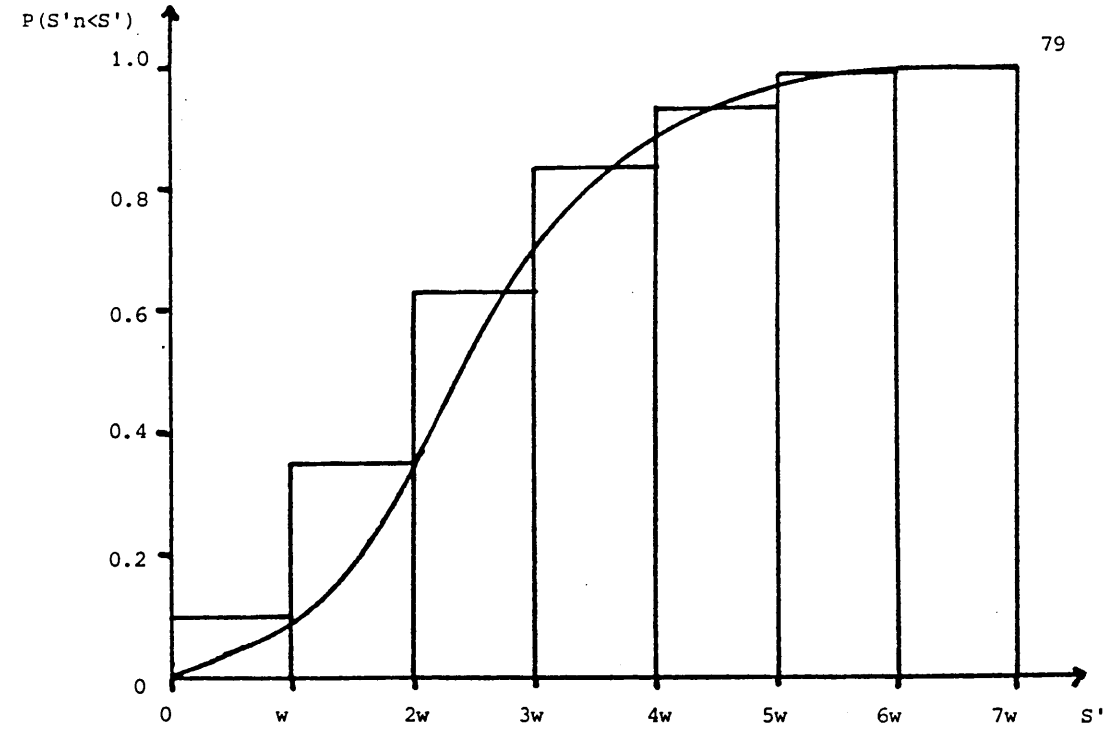


Figure 14a. Cumulative distribution of Sn and S'n where n=5, r=2 and a=0

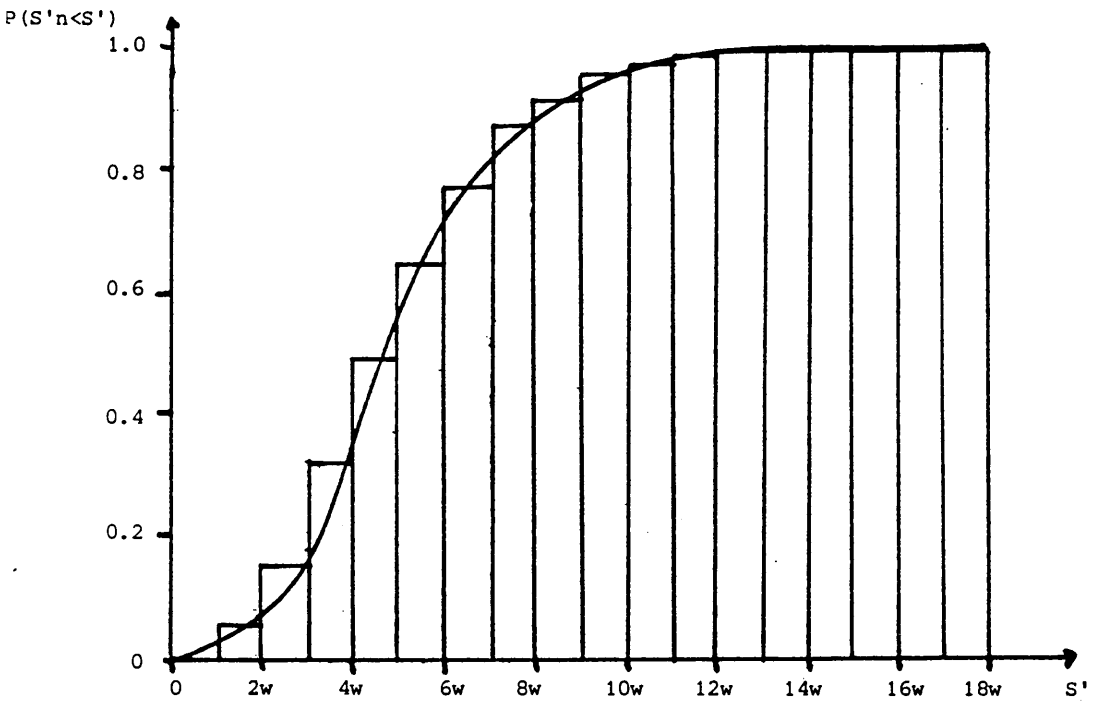


Figure 14b. Cumulative distribution of Sn and S'n where n=5, r=1 and a=0

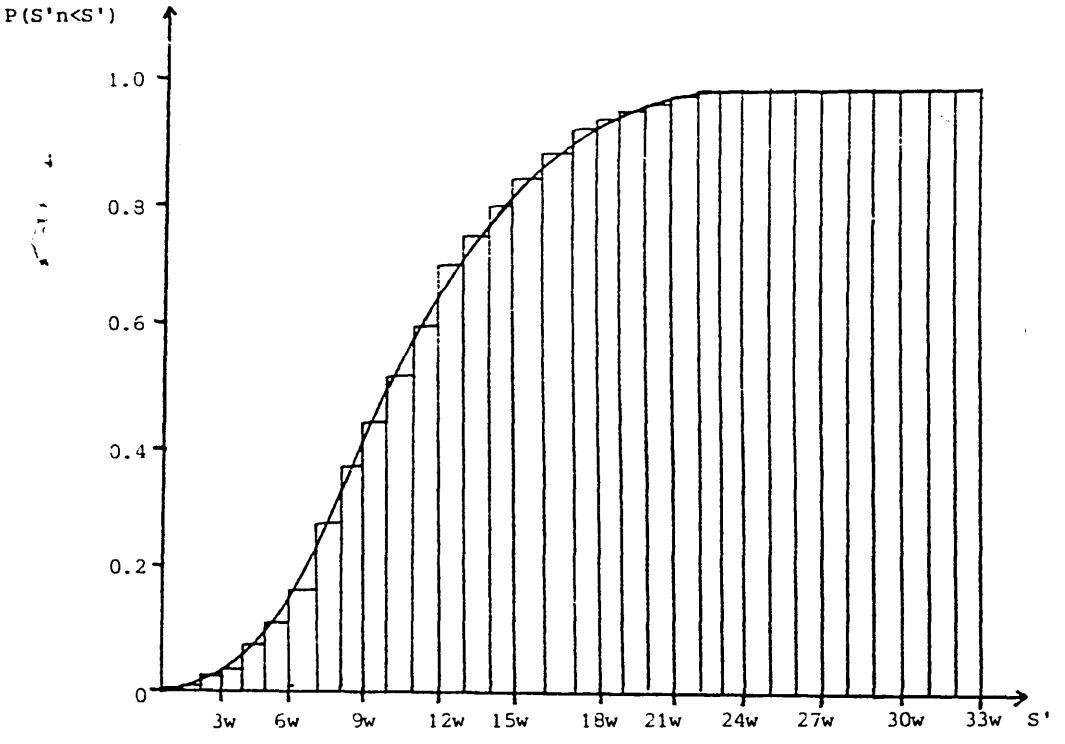


Figure 14c. Cumulative distribution of Sn and S'n where n=5, r=0.5 and a=0

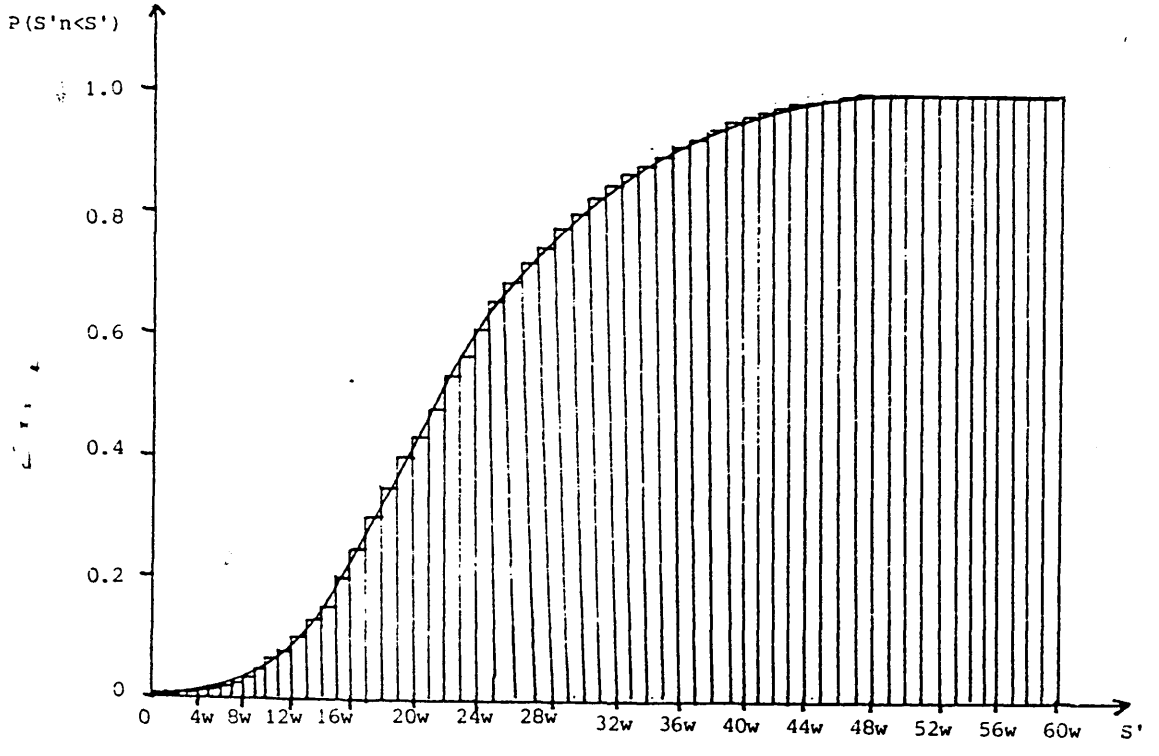


Figure 14d. Cumulative distribution of Sn and S'n where n=5, r=0.25 and a=0

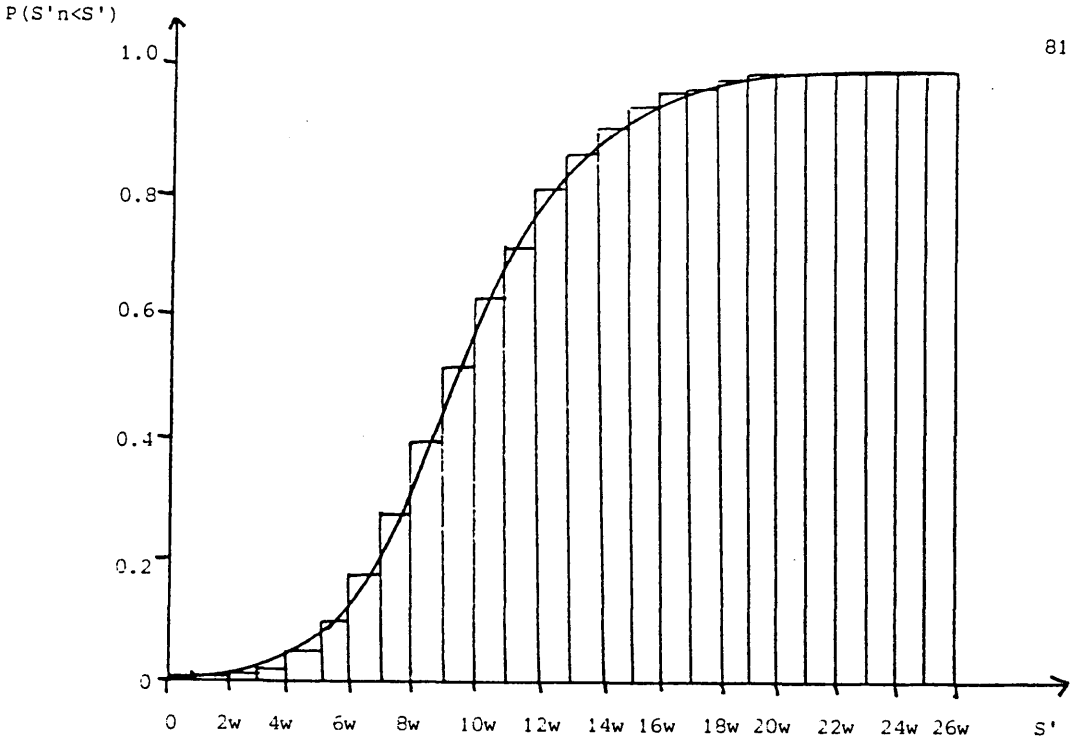


Figure 14e. Cumulative distribution of Sn and S'n where n=10, r=1 and a=0

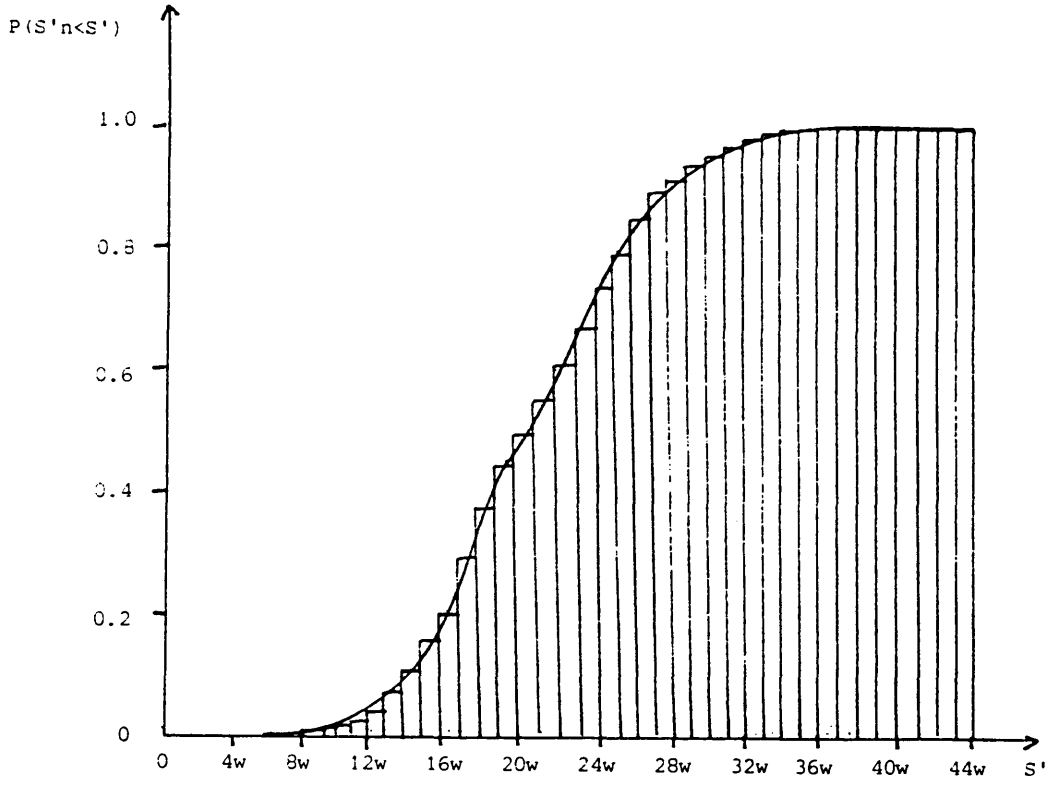


Figure 14f. Cumulative distribution of Sn and S'n where n=20, r=1 and a=0

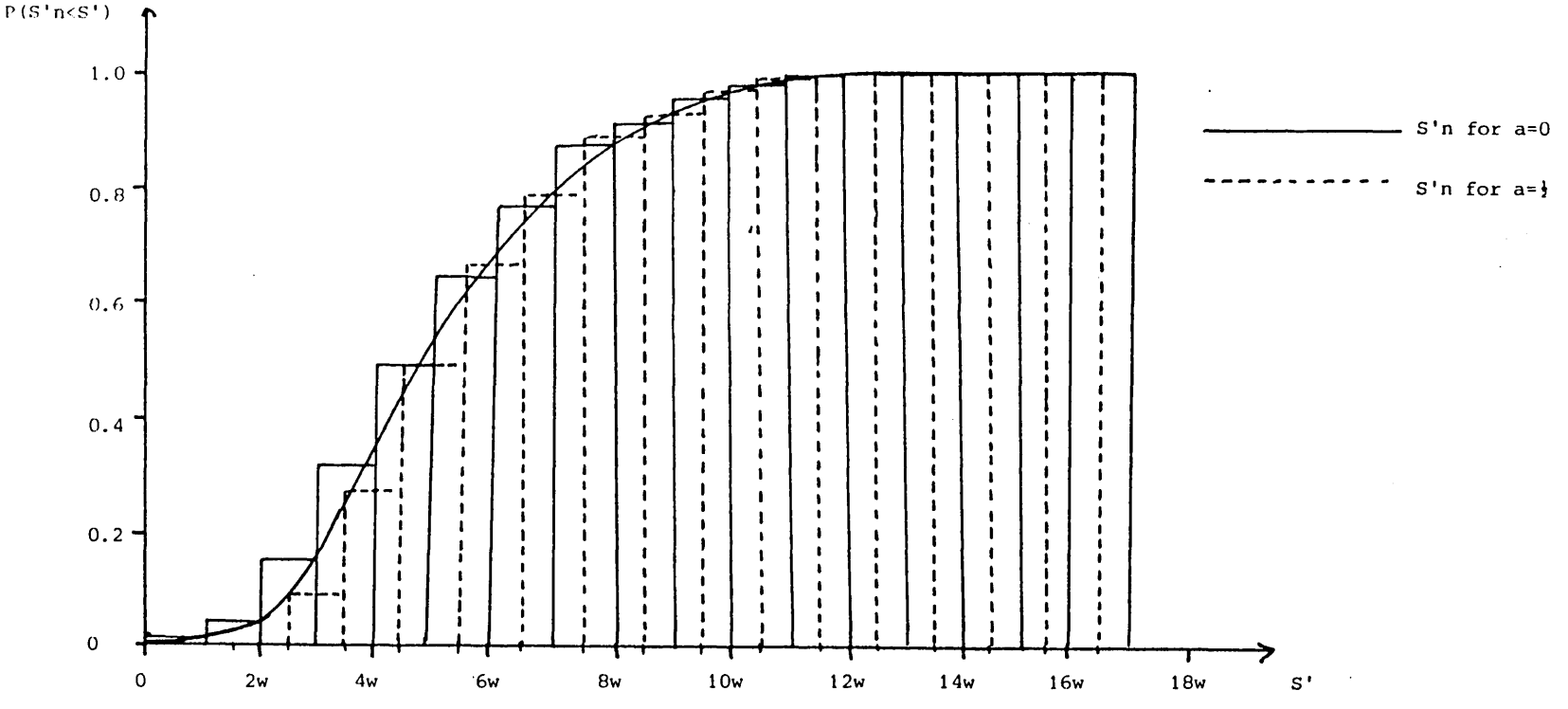


Figure 14g. Cumulative distribution of Sn and S'n where n=5, r=1, a=0 and a=½

TABLE 1 Values of the Function h(z)

Z	h(z)				
	r=2	r=1	r=0.5	r=0.25	r=0.1
-0.5w	-14.91	-4.05	-1.03	-0.26	-0.04
-0.45w	-5.96	0.87	1.47	0.99	0.46
-0.40w	3.93	6.04	4.04	2.27	0.96
-0.35w	14.86	11.48	6.67	3.55	1.47
-0.30w	26.94	17.20	9.37	4.85	1.98
-0.25w	40.29	23.20	12.14	6.17	2.49
-0.20w	55.05	29.52	14.98	7.50	3.00
-0.15w	71.35	36.16	17.89	8.86	3.52
-0.10w	89.38	43.14	20.88	10.23	4.04
-0.05w	109.29	50.33	23.94	11.61	4.56
0.00w	-68.70	-41.80	-22.92	-11.98	-4.92
0.05w	-65.40	-38.81	-20.97	-10.87	-4.44
0.10w	-61.77	-35.68	-18.97	-9.75	-3.96
0.15w	-57.74	-32.38	-16.92	-8.62	-3.48
0.20w	-53.30	-28.92	-14.82	-7.47	-3.00
0.25w	-48.40	-25.27	-12.66	-6.30	-2.57
0.30w	-42.96	-21.44	-10.45	-5.12	-2.02
0.35w	-36.96	-17.41	-9.35	-3.93	-1.53
0.40w	-30.33	-13.18	-5.86	-2.72	-1.04
0.45w	-23.01	-8.73	-3.28	-1.50	-0.54
0.50w	-14.91	-4.05	-1.03	-0.26	-0.04

TABLE 2 Endpoints of the possible range in values of E(Z) for values of r

r	2.0	1.0	0.5	0.25	0.1
Range of E[Z]	31.40 to -16.10(1)	8.20 to -4.10	2.07 to -1.03	0.52 to -0.26	0.08 to -0.04

Range of E[Z] expressed as a percentage of θ

(1) indicates E[Z] lies between 31.40 and -16.10 θ

TABLE 3 Maximum errors expected for mean and variance caused by rounding

r	Max error in mean (% of standard deviation)	Max error in variance (% of variance)
0.10	0.1 θ	0.2 θ^2
0.25	0.5 θ	1.1 θ^2
0.50	2.1 θ	4.0 θ^2
1.00	8.2 θ	17.5 θ^2
1.50	18.1 θ	40.0 θ^2
2.00	31.4 θ	72.2 θ^2

TABLE 4 Values of a which make $\bar{X}'$ an unbiased estimate of θ

r	Value of a for which $E[X'] = \theta$	
2.00	0.233	-0.340
1.50	0.247	-0.328
1.00	0.261	-0.316
0.75	0.268	-0.309
0.50	0.275	-0.302
0.33	0.279	-0.298
0.25	0.282	-0.296
0.10	0.286	-0.291

TABLE 5 Examples of Compensation Calculations

r	n	X'	$\hat{\theta}$	$\lvert X' - \hat{\theta} \rvert$	$\lvert \hat{\theta} - \hat{\theta} \rvert$
0.5	50	0.860	0.872	0.140	0.128
	100	0.840	0.852	0.160	0.148
	250	1.102	1.111	0.102	0.111
	500	1.070	1.080	0.070	0.080
	1000	0.966	0.976	0.034	0.024
1.0	50	0.780	0.828	0.220	0.172
	100	0.953	0.994	0.047	0.006
	250	1.032	1.070	0.032	0.070
	500	0.933	0.975	0.067	0.025
	1000	0.922	0.964	0.078	0.036
1.5	50	0.750	0.851	0.250	0.149
	100	0.780	0.878	0.220	0.122
	250	0.910	0.998	0.090	0.002
	500	0.847	0.940	0.153	0.060
	1000	0.905	0.993	0.095	0.007
2.0	50	0.760	0.919	0.240	0.081
	100	0.853	1.002	0.147	0.002
	250	0.904	1.048	0.096	0.048
	500	0.836	0.986	0.164	0.014
	1000	0.922	1.064	0.078	0.064
3.0	50	0.420	0.756	0.580	0.244
	100	0.680	0.979	0.320	0.021
	250	0.540	1.117	0.160	0.117
	500	0.756	1.044	0.244	0.044
	1000	0.670	0.970	0.330	0.030
7.0	50	0.280	1.087	0.720	0.087
	100	0.210	0.998	0.790	0.002
	250	0.175	0.949	0.825	0.051
	500	0.252	1.052	0.748	0.052
	1000	0.214	1.003	0.786	0.003

TABLE 6 Range of α_1' values

	r					
	2.0	1.5	1.0	0.5	0.25	0.1
n=5 α_1 = 0.001	0-0.117	0-0.049	0-0.029	0-0.004	0-0.002	0.001-0.001
0.010	0-0.229	0-0.075	0-0.055	0.006-0.028	0.007-0.015	0.009-0.011
0.050	0-0.321	0-0.183	0-0.102	0.028-0.065	0.041-0.056	0.048-0.053
n=15 α_1 = 0.001	0-0.0084	0-0.034	0-0.011	0.001-0.003	0.001-0.001	0.001-0.001
0.010	0-0.220	0-0.078	0-0.033	0.006-0.015	0.009-0.011	0.010-0.010
0.050	0-0.313	0-0.180	0.007-0.111	0.029-0.062	0.044-0.053	0.048-0.051
n=25 α_1 = 0.001	0-0.081	0-0.030	0-0.008	0.001-0.002	0.001-0.001	0.001-0.001
0.010	0-0.152	0-0.087	0.001-0.037	0.006-0.015	0.008-0.011	0.010-0.010
0.050	0-0.331	0-0.166	0.006-0.107	0.038-0.065	0.048-0.056	0.049-0.051